Model Building for Semiparametric Mixtures

Ramani S. Pilla, Francesco Bartolucci and Bruce G. Lindsay[1]

Abstract

An important and yet difficult problem in fitting multivariate mixture models is determining the mixture complexity. We develop theory and a unified framework for finding the nonparametric maximum likelihood estimator of a multivariate mixing distribution and consequently estimating the mixture complexity. Multivariate mixtures provide a flexible approach to fitting high-dimensional data while offering data reduction through the number, location and shape of the component densities. The central principle of our method is to cast the mixture maximization problem in the concave optimization framework with finitely many linear inequality constraints and turn it into an unconstrained problem using a *penalty function*. We establish the existence of parameter estimators and prove the convergence properties of the proposed algorithms. The role of a "sieve parameter" in reducing the dimensionality of mixture models is demonstrated. We derive analytical machinery for building a collection of semiparametric mixture models, including the multivariate case, via the sieve parameter. The performance of the methods are shown with applications to several data sets including the cdc15 cell-cycle yeast microarray data.

Key Words: Data reduction; High-dimensional modeling; Multivariate normal distribution; Nonparametric maximum likelihood; Nonparametric density estimation; Penalty function.

[1] Pilla is with Department of Statistics, Case Western Reserve University, Cleveland, OH, USA (E-mail: pilla@case.edu). Bartolucci is with Department of Economics, University of Urbino, Urbino, Italy (E-mail: Francesco.Bartolucci@uniurb.it). Lindsay is with Department of Statistics, Pennsylvania State University, University Park, PA, USA (E-mail: bgl@psu.edu). Pilla's research was partially supported by the National Science Foundation (NSF) grant DMS 02-39053 and Office of Naval Research grants N00014-02-1-0316 & N00014-04-1-0481. Bartolucci's research was partially supported by the MIUR grant 2002. Lindsay's research was partially supported by the NSF grants DMS 01-04443 and DMS 04-05637.

1 Introduction

Multivariate mixture modeling is a bridge between clustering and nonparametric multivariate density estimation. The estimated multivariate mixture model provides both an estimate of the density for the overall data and partitions the data into several components or clusters. Determining the mixture complexity is challenging even in one dimension (Laird, 1978; Jewell, 1982; Titterington et al., 1985; Lesperance and Kalbfleisch, 1992; Roeder, 1994; Lindsay, 1995; McLachlan and Peel, 2001; Pilla and Loader, 2003; Scott, 2004a; Pilla and Charnigo, 2005). A fundamental problem in high-dimensional modeling is determining the number of components and their centers. One popular model-free approach to high-dimensional modeling is the *K-means* algorithm (see Hastie et al. (2001) and the references therein). Model-based techniques such as density estimation (Scott, 1992; James et al., 2001; Scott, 2004b) and multivariate mixture models provide a reliable and flexible approach to high-dimensional modeling while providing a data reduction through the number, location and shape of the component densities. In the context of mixture models, the problem becomes determining the number of mixture components and estimating the corresponding location parameter vectors. Furthermore, mixture models provide much of the flexibility of the nonparametric approaches, while retaining many advantages of the parametric approaches (Laird, 1978; Roeder, 1992; Lesperance and Kalbfleisch, 1992; Lindsay, 1995; Charnigo and Pilla, 2005; Scott, 2004b).

One of the main reasons for the popularity of model-free approaches such as the K-means algorithm for high-dimensional modeling is the lack of a unified and powerful technique for fitting multivariate mixtures. The focus of this article is to develop analytical machinery for *building a collection of semiparametric mixture models*, including the multivariate case. The theory and methods developed in this article have applications to image analysis, high-dimensional clustering and data mining, to name a few. The practical applications of semiparametric mixture models are broad and include case-control studies with errors-in-variables (Roeder et al., 1996), random effects models and empirical Bayes method (Lindsay, 1995). A natural outcome of applying a multivariate mixture model for high-dimensional clustering is that (1) each cluster is statistically represented by a parametric distribution; for instance, normal in continuous case and Poisson in discrete case and (2) it provides the proportion of observations in each cluster through the estimated mixture probability. Furthermore, statistical tests can be developed easily based on the parameter estimators of the multivariate mixture models to answer various scientific or biological questions. Due to high levels of noise inherent in many of the massive data sets, including the microarray technology, it is highly desirable to carry out the high-dimensional data analysis within a statistical framework.

Let $m := |\text{supp}(\mathcal{Q})|$ be the size of the support set of a *mixing measure* $\mathcal{Q}$; i.e., the *mixture complexity*. If $\mathcal{Q}$ is finitely supported, m is finite and otherwise it is infinite. The current standard approach for finding the maximum likelihood estimator (MLE) of $\mathcal{Q}$, when m is known a priori or fixed, is the well-known EM algorithm developed in the seminal article by Dempster et al. (1977).

In the absence of the knowledge of mixture complexity, it is instructive to start with an *overparameterized mixture model* and search over the whole continuous parameter space effectively to obtain a parsimonious mixture model. Overparameterization here refers to fitting a model with many components relative to the actual number in the nonparametric MLE (NPMLE) of $\mathcal{Q}$; hence, there is a redundancy of components in the mixture model. Such a scheme would be robust to parameter starting values chosen for fitting the mixture model. To accomplish this, one requires a powerful mixture algorithm that pushes most of the mixture probabilities to zero, which is on the boundary of the parameter space. The focus of this article is to develop theory and create robust (to starting values) as well as powerful algorithms to address this problem.

The popular and widely employed EM-based algorithms are particularly slow for fitting such an overparameterized mixture model and it is very difficult, if not impossible, to remove the unnecessary components; since the algorithm can never reach such a boundary point. Moreover, the EM algorithm is sensitive to parameter starting values and fails to converge in certain mixture problems. Figure 4 in Section 7.1 demonstrates this aspect of the algorithm. Other examples where the EM algorithm converges to saddle points or fails to converge are noted by McLachlan and Krishnan (1997).

1.1 Statistical Framework

Let $\mathcal{F} := \{f_{\boldsymbol{\theta}}(\mathbf{x}) : \boldsymbol{\theta} \in \Omega \subset \Re^p\}$ be a family of probability density functions with respect to a σ-finite dominating measure μ for a p-dimensional random vector $\mathbf{x} \in \mathcal{X} \subset \Re^n$ and a p-dimensional location parameter vector $\boldsymbol{\theta} \in \Omega \subset \Re^p$, a measurable space. Assume that the component density $f_{\boldsymbol{\theta}}(\mathbf{x})$ is bounded in $\boldsymbol{\theta}$ for each $\mathbf{x} \in \mathcal{X}$. Let $\mathcal{G}$ be the space of all probability measures on Ω with the σ-field generated by its Borel subsets. For a given $\mathcal{Q} \in \mathcal{G}$, we assume that data vector $\mathbf{x}$ arises from the marginal density

$$\mathrm{g}_{\mathcal{Q}}(\mathbf{x}) = \int f_{\boldsymbol{\theta}}(\mathbf{x})\, d\mathcal{Q}(\boldsymbol{\theta}) \quad \text{for} \quad \mathbf{x} \in \mathcal{X} \tag{1}$$

which is referred to as a *mixture density*. The mixture model (1) is also applicable to empirical Bayes estimation, where $\mathcal{Q}$ is an unknown prior distribution and the objective becomes estimation of the posterior distribution of $\boldsymbol{\theta}$ without assuming a functional form for the prior distribution.

The goal is to estimate the *mixing measure* $\mathcal{Q}$ by finding the probability measure $\widehat{\mathcal{Q}} \in \mathcal{G}$ that maximizes the nonparametric mixture loglikelihood $\log\{L(\mathcal{Q})\} = \sum_{i=1}^{n} \log\{\mathrm{g}_{\mathcal{Q}}(\mathbf{x}_i)\}$. It is well known that finding the NPMLE of $\mathcal{Q}$ is computationally intensive (Lesperance and Kalbfleisch, 1992; Roeder, 1992; Lindsay, 1995; Bickel et al., 1998; Susko et al., 1999). Although $\mathcal{Q}$ is an arbitrary probability measure, under mild conditions Lindsay (1983a,b) showed that finding the MLE involved a standard problem of convex optimization, that of maximizing a concave function over a convex set. One consequence of this is that, as long as $l(\mathcal{Q})$ is bounded, the MLE of $\mathcal{Q}$ is concentrated on a support of cardinality at most that of d—the *number of distinct observed data vectors*. This is a very useful, albeit surprising, result since a potentially difficult nonparametric estimation problem is reduced to that of a finite dimension; hence algorithms can be constructed to find the solution. Hence, we restrict the attention to discrete probability measures $\mathcal{Q}$ having p-dimensional *support vectors* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ collected in a matrix $\boldsymbol{\Theta}$ with a corresponding vector of *masses* or *mixing probabilities* denoted by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)^T$ such that $\boldsymbol{\pi}$ is in the *unit simplex* $\boldsymbol{\Pi} := \{\boldsymbol{\pi} \in \Re^m : \pi_j \in [0,1], \sum_{j=1}^{m} \pi_j = 1\}$.

We consider model fitting for both discrete and continuous data. Therefore, it is instructive to define $\mathbf{f}_{\boldsymbol{\theta}} := \{f_{\boldsymbol{\theta}}(\mathbf{y}_1), \ldots, f_{\boldsymbol{\theta}}(\mathbf{y}_d)\}^T$ to be the d-dimensional vector of distinct likelihood terms, where T denotes transpose, $(\mathbf{y}_1, \ldots, \mathbf{y}_d) \in \mathcal{Y}$ are the distinct observation vectors arising from the original data vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}$. Let n_i be the number of times $\mathbf{y}_i$ occurs in the sample of $\mathbf{x}$ vectors. For continuous data, $n = d$ and for discrete data, often $n \gg d$.

The observed data matrix of dimension $(n \times p)$ with row vectors $\mathbf{y}_i$ $(i = 1, \ldots, n)$ is assumed to arise from the mixture density $\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i) = \sum_{j=1}^{m} \pi_j \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)$, and the discrete mixing measure can be represented as $\mathcal{Q} = \sum_j \pi_j \, \varrho(\boldsymbol{\theta}_j)$, where $\varrho(\boldsymbol{\theta})$ is a discrete measure with mass one at $\boldsymbol{\theta} \in \Omega$. If m is fixed, the model $\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)$ will be referred to as the *m-component mixture model*, and one can always find the NPMLE of $\mathcal{Q}$ using m equal to d (Lindsay, 1983a,b). However, the actual number of distinct support vectors with positive mixture probability, referred to as *active supports*, can be as small as one. The mixture loglikelihood of $\mathcal{Q}$ becomes

$$l(\mathcal{Q}) = \log L(\mathcal{Q}) = \sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)\} \quad \text{over all} \quad \mathcal{Q} \in \mathcal{G}. \tag{2}$$

The goal is to find $\widehat{\mathcal{Q}} \in \mathcal{G}$ such that $l(\widehat{\mathcal{Q}}) = \sup_{\mathcal{Q} \in \mathcal{G}} l(\mathcal{Q})$.

The biggest practical problem one faces in solving the loglikelihood equations in (2) is that the number of inequalities is equal to the number of elements of the parameter space Ω. There are some important problems where this number is finite, although possibly very large, such as in target recognition, hyperspectral image analysis and positron emission tomography. For these problems discretization of the parameter space is directly relevant. In other problems, Ω may be a continuous space; hence, one needs a machinery for approximating the parameter space to solve these equations (see Section 2.1).

1.2 Main Results

In this article, we develop a unified framework for finding the NPMLE of a multivariate mixing distribution $\mathcal{Q}$ and consequently for building a collection of semiparametric mixture models. The key ingredients for building these models are the "sieve parameter" controlling the dimensionality of the mixture problem (as shown in Figures 2 and 3) and the ability to fit overparameterized mixture models. This collection of models enable us to investigate the role of many overlapping densities, thereby creating an ideal situation for solving large-scale practical problems.

We create a powerful technique referred to as the "Penalized Dual method" and an efficient algorithm for fitting overparameterized mixture models. Consequently we have a method for estimating the mixture complexity. This algorithm is a step in the direction of developing a unified framework for building a collection of semiparametric mixture models.

The underlying principle for our method is that the mixture loglikelihood forms a concave functional on the convex set of all probability distributions which implies that there exists a *dual optimization problem* [Section 5.3, Lindsay (1995)]. Lesperance and Kalbfleisch (1992) and Susko et al. (1999) exploited this to create an elegant algorithm for finding the NPMLE of $\mathcal{Q}$ in univariate mixtures. This research is in the same spirit but extends these ideas by introducing a "penalty term". A fundamental feature of our approach is that it eliminates ad hoc procedures to estimate the penalty parameter. The dual problem has a statistical interpretation analogous to the least squares problem and the formulation is strikingly similar to the one that arises in empirical likelihood (Owen, 2001) framework (see Section 3.1).

Our main results are summarized as follows.

1. In Section 2, we cast the mixture problem in the dual optimization framework. We first develop a machinery for approximating the continuous parameter space Ω and next create an algorithm (based on the Penalized Dual method) for finding the maximum of $l(\mathcal{Q})$. Consequently, we propose an algorithm for estimating the mixture complexity. We establish the convergence of this algorithm to MLE in Section 5.
2. In Section 3, we develop theory for the Penalized Dual method in solving the dual optimization problem while presenting the statistical interpretation for our framework. By exploiting the inherent advantage of the penalty formulation, we derive a technique for converting parameter estimators from the Penalized Dual problem into the mixture probability parameters. We show that the Penalized Dual estimators converge to the mixture probability estimators as the penalty is increased, with the correct limits.
3. Section 4 establishes the existence of parameter estimators and derives convergence results for the Penalized Dual algorithm, for fitting overparameterized mixture models.

The Penalized Dual algorithm effectively yields an estimate for the mixture complexity. Our algorithm is based on a modification of the Newton-Raphson algorithm and therefore, it inherits its virtues while retaining the stability (i.e., monotonically increasing the likelihood) of EM-based algorithms. Empirical assessment of the faster rate of convergence of our stable and powerful algorithm compared to the EM algorithm will be demonstrated in Section 5.

4. It is shown that the algorithm based on the Penalized Dual method is robust to choice of parameter starting values and achieves the global maximum. The dimension of the dual optimization problem is fixed at d, the number of distinct observed data vectors, whereas for the mixture problem it grows with the mixture complexity m. For discrete mixture problems, such as binomial or Poisson, d can be much smaller than n (see for example Section 7.1). In these cases, the Penalized Dual method has no dimensionality cost.

5. In Section 6, we derive several important structural properties of multivariate normal mixtures in which $\mathcal{Q}$ is modeled nonparametrically in the presence of an unknown variance-covariance matrix $\mathbf{\Sigma} \in \mathcal{S}$ common to all m components, where $\mathcal{S}$ is a compact space. The power of our method rests in building a collection of semiparametric mixture models, including the multivariate case. We demonstrate the role of the sieve parameter in reducing the dimension of the mixture problem by creating novel graphical devices (Figures 2 and 3) referred to as *Mixture Tree Plots*.

6. When the cardinality of the discrete parameter set (chosen for approximating the continuous parameter space Ω) is large, the EM algorithm for such a mixture problem fails to converge to the MLE, for all practical purposes, while the Penalized Dual algorithm converges. From a model selection point of view, the EM algorithm does not eliminate the redundant components while the Penalized Dual algorithm yields a parsimonious mixture model. Empirical evidence of this is shown in Figure 4, Section 7.1.

7. Section 7 illustrates the power of the proposed methods using several applications. We compare our method with the EM algorithm, due to lack of a unified and/or stable algorithms for fitting the overparameterized mixture problems and for building

semiparametric mixture models. For the univariate mixture case, we compared with Rotated EM algorithm (an accelerated version of the EM only applicable to the univariate mixtures) proposed by Pilla and Lindsay (2001). Our empirical investigation demonstrate the faster rate of convergence of our algorithm, compared with the EM algorithm.

Section 8 presents the conclusions and the Appendix derives technical details.

1.3 Relevant Literature

Widely employed model-free methods for high-dimensional modeling include the K-means algorithm, hierarchical clustering and agglomerative and divisive algorithms (Hastie et al., 2001). However, none of these techniques take advantage of the inherent statistical structure of the data.

The existing model-based mixture algorithms include those for finding the NPMLE of $\mathcal{Q}$ (Lesperance and Kalbfleisch, 1992; Susko et al., 1999; Connolly et al., 2001). These algorithms are either not fast enough for high-dimensional modeling or not applicable for the following mixture problem: (1) the component densities are poorly separated and/or (2) many of the estimated mixture probabilities are on the boundary of the parameter space. In analyzing Sloan Digital Sky Survey data by fitting the multivariate normal mixtures, Connolly et al. (2001) noted that many existing techniques are not computationally efficient and their mixture EM algorithm obtains an improvement of only three orders of magnitude. Therefore, developing a powerful method for fitting multivariate mixtures is desirable.

Although there have been some promising developments on accelerating the EM algorithm (see McLachlan and Krishnan (1997) and the references therein), none of these methods address the overparameterized mixture problem described earlier. To overcome the above difficulties, Pilla and Lindsay (1996, 2001) proposed alternative augmentation schemes based on the principles of the EM that provide a significantly improved convergence rate of the EM algorithm for a class of finite mixture models. At this time, it is not clear how to extend these methods to multivariate mixture models; however, they do provide an important class for comparison with our algorithm in univariate mixture problems (comparisons are made in Section 7.1). Lindsay (1995, Section 6.3) discusses several algorithmic methods based on directional derivatives such as the vertex direction method and vertex exchange method to find the NPMLE of $\mathcal{Q}$. These methods also require searching over a discrete parameter space and have certain computational disadvantages (Lesperance and Kalbfleisch, 1992).

2 Mixture Maximum Likelihood Problems

In this section we first formulate the mixture problem as a convex optimization problem and next create a framework for approximating the continuous parameter space. Lastly, we develop an algorithm for finding the NPMLE of $\mathcal{Q}$. This algorithm forms the basis for building a collection of semiparametric mixture models developed in Section 6.

2.1 Maximizing $l(\mathcal{Q})$ via Approximating $\mathcal{G}$

If the number of components in $\mathcal{Q}$ is fixed, but the location parameter vectors are unknown, then $l(\mathcal{Q})$ can have several local maxima (Lesperance and Kalbfleisch, 1992; Lindsay, 1995; Pilla and Lindsay, 2001). Both the EM and the K-means algorithms can get trapped at a local maximum while requiring a priori knowledge of the mixture complexity m. To overcome this problem, researchers often randomly perturb the parameter starting values and recompute the local maxima (Hall and Zhou, 2003; Hunter, 2004). However, there is no theoretical justification to guarantee that the resulting solution reaches closer to the global maximum.

In fact, random parameter starting values can fail in the mixture context for the following reasons. First, one requires an a priori knowledge of the number of components m. Second, there is a danger of choosing multiple starting values from one component while ignoring to choose any from other components. In such a case, the EM algorithm may not necessarily be able to locate the component from which no parameter values are selected. This problem becomes severe when components of unequal sizes are present; see Section 7.3 for an empirical investigation of this aspect.

To combat the difficulties with the parameter starting value problem and an a priori knowledge of m, we develop a technique in which we approximate $\mathcal{G}$ by the set of all probability measures on a discrete parameter space of Ω.

Approximating $\mathcal{G}$: Approximate $\mathcal{G}$ by $\mathcal{G}_m$, where $\mathcal{G}_m$ is a set of discrete distributions generated by a finite subset of Ω. We set this finite subset to be $\boldsymbol{\Theta}_m = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$. As $m \to \infty$ and $\boldsymbol{\Theta}_m$ becomes dense in Ω, the set $\mathcal{G}_m \to \mathcal{G}$. In practice, a sufficiently large m is chosen such that $\mathcal{G}_m$ approximates $\mathcal{G}$ well. Therefore, the cardinality of $\boldsymbol{\Theta}_m$, namely m, determines how close the MLE is to the global MLE over all measures on Ω. In approximating $\mathcal{G}$, it is important to select a suitable $\boldsymbol{\Theta}_m$ while keeping computations manageable. This will be addressed in Section 6.3.

In what follows, we distinguish between the three mixture problems.

1. The *fixed support mixture* problem is equivalent to maximizing

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{d} n_i \log \left\{ \mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i) \right\} \tag{3}$$

   over the parameter space $\boldsymbol{\Pi}$ while treating the support set $\boldsymbol{\Theta}_m \subset \Omega$ as fixed. This is the *primal* or *mixture problem* for which we define a "dual" in Section 3. Note that $\dim(\boldsymbol{\pi}) = (m-1)$ and grows with the cardinality of $\boldsymbol{\Theta}_m$, which is a major obstacle when $\dim(\boldsymbol{\Theta}_m)$ is large. *However, the dimension of our dual optimization problem is fixed at d, the number of distinct observed data vectors.*

2. We fix the number of components in the mixing distribution $\mathcal{Q}$ to be m but treat the $\boldsymbol{\theta}$ parameter vectors as unknown for each component. Therefore, the *continuous support mixture model* problem becomes simultaneously estimating $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ parameter vectors by maximizing $l(\mathcal{Q})$ over $\boldsymbol{\Pi} \times \boldsymbol{\Theta}_m$ for a fixed m.

3. In the absence of knowledge of mixture complexity m, maximizing the mixture loglikelihood in (2) yields an NPMLE that is a discrete distribution on the parameter space with a random number of component densities (Lindsay, 1995; Pilla and Lindsay, 2001). This will be referred to as the *nonparametric mixture model*. The goal in turn becomes finding the probability measure $\widehat{\mathcal{Q}} \in \mathcal{G}$ that maximizes (2).

For discrete mixture problems, such as binomial or Poisson, often $d \ll n$; therefore, the dual methods are able to reduce the dimension of the mixture problem. The effect of this dimensionality on the performance of the algorithms will be demonstrated in Section 7.1.

2.2 Characterization of the NPMLE of $\mathcal{Q}$

Let Γ be a curve in $\Re^d$ consisting of all vectors of the form $\{f_{\boldsymbol{\theta}}(\mathbf{y}_1), \ldots, f_{\boldsymbol{\theta}}(\mathbf{y}_d)\}$, where $\boldsymbol{\theta} \in \Omega$. Under compactness of Γ, we can define the convex hull of Γ as $\text{Conv}(\Gamma) = \{\mathbf{g}_{\mathcal{Q}} : \mathcal{Q} \in \mathcal{G},\ \mathcal{Q} \text{ has finite support}\}$, where $\mathbf{g}_{\mathcal{Q}} = \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_1), \ldots, \mathrm{g}_{\mathcal{Q}}(\mathbf{y}_d)\}^T$. The optimal vector $\mathbf{g}_{\widehat{\mathcal{Q}}} = \{\mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_1), \ldots, \mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_d)\}^T \in \text{Conv}(\Gamma)$ and a corresponding maximizing measure $\widehat{\mathcal{Q}}$ can be characterized in terms of the gradient function as shown next.

Definition 1 (Finite identifiability): For a given family $\mathcal{F}$, suppose that $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{G}$ have finite support. Suppose that $\mathcal{Q}_j \in \mathcal{G}$ yields the mixture density $\mathrm{g}_{\mathcal{Q}_j}(\mathbf{y})$ for $j = 1, 2$. If $\mathrm{g}_{\mathcal{Q}_1}(\mathbf{y}) = \mathrm{g}_{\mathcal{Q}_2}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$ implies $\mathcal{Q}_1 = \mathcal{Q}_2$, then the corresponding collection of mixture densities is said to have the *finite identifiability* property.

An important aspect of our technique is based on the following fundamental property. For the NPMLE $\widehat{\mathcal{Q}}$, the ith *fitted model* $\mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i)$ is guaranteed to be unique (regardless of identifiability of the mixture density), and that one can determine these fitted values by solving for the *residual* $\widehat{w}_i$, on a log-scale, defined as

$$\log\,(\widehat{w}_i) := \log\left(\frac{n_i}{n}\right) - \log\{\mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i)\} \quad \text{for} \quad \mathbf{y}_i \in \mathcal{Y},\, i = 1, \ldots, d. \tag{4}$$

In ordinary parametric likelihood problems the solution is characterized by the likelihood equations. We extend these ideas to our problem to show that the fitted values and the corresponding mixing distribution $\mathcal{Q} \in \mathcal{G}$ can be further characterized in terms of a set of gradient equations. That is, $\widehat{\mathcal{Q}}$ is an NPMLE if and only if

$$\Psi\left(\widehat{\mathcal{Q}}\right) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_m} \mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}) \leq 0 \quad \text{for} \quad \mathcal{Q} \in \mathcal{G}, \tag{5}$$

where the *gradient function*, the directional derivative of the mixture loglikelihood in the direction of a component density, is defined as

$$\mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}) := \sum_{i=1}^{d} n_i \left\{ \frac{f_{\boldsymbol{\theta}}(\mathbf{y}_i)}{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)} - 1 \right\} \quad \text{for} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_m. \tag{6}$$

If a candidate maximizing measure $\widehat{\mathcal{Q}}$ violates the *gradient inequality* in (5) at some $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$, then one is not at the maximum. In particular, one can increase the loglikelihood by placing some positive probability at $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$.

2.3 Finding the Maximum of $l(\mathcal{Q})$ via $\boldsymbol{\Theta}_m$

For a fixed m, mixture estimation is challenging due to the fact that $l(\mathcal{Q})$ is not concave and hence there are several local maxima (Lesperance and Kalbfleisch, 1992; Lindsay, 1995; McLachlan and Peel, 2001; Pilla and Lindsay, 2001). We create an algorithm that is robust to the choice of parameter starting values and reaches closer to the global maximum of $l(\mathcal{Q})$.

We find the NPMLE of $\mathcal{Q}$ adaptively as follows.

Algorithm 1 [Finding the Maximum of $l(\mathcal{Q})$]

1. Consider $\boldsymbol{\Theta}_m \subset \Omega$ to be the support set of $\mathcal{Q}$. Solve the fixed support mixture problem by maximizing (3) over the parameter space $\boldsymbol{\Pi}$ on the support set $\boldsymbol{\Theta}_m$, while treating $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$ as fixed. It is worth noting that the larger the cardinality of $\boldsymbol{\Theta}_m$, the higher the value of the loglikelihood at convergence.
2. Apply the MLE $\widehat{\boldsymbol{\pi}}$ (with the corresponding fixed support set $\boldsymbol{\Theta}_m \subset \Omega$) obtained in Step 1, as parameter starting values for the continuous support mixture problem and maximize (2) over the product parameter space $\boldsymbol{\Pi} \times \boldsymbol{\Theta}_m$.

For Step 1, one requires a stable and powerful mixture algorithm and is derived in the next sections. In particular, Algorithm 2 presented in Section 4.2 can be employed in Step 1. The Step 2 may include estimation of other parameters in the model such as $\boldsymbol{\Sigma} \in \mathcal{S}$, in the multivariate normal mixtures context.

For the continuous support mixture model, it will be shown in Section 7 that Algorithm 1 reaches closer to the global maximum, if not to the global maximum. Our empirical evidence suggests that Algorithm 1 is superior to EM-type algorithms that start with random (or arbitrary) parameter values.

3 The Dual Optimization Problem: Properties of Estimators

We now present the problem that is dual to the primal problem (3) considered by Lindsay (1983a) and develop theory for effectively solving it. The *dual problem* is to maximize

$$l(\mathbf{w}) = \sum_{i=1}^{d} n_i \log(w_i) \tag{7}$$

subject to the constraints $\mathbf{w} = (w_1, \ldots, w_d)^T \in \Re_+^d$ and

$$\sum_{i=1}^{d} w_i\, f_{\boldsymbol{\theta}}(\mathbf{y}_i) \leq 1 \quad \text{for} \quad \boldsymbol{\theta} \in \Omega. \tag{8}$$

Let $\widehat{\mathbf{w}} \in \Re_+^d$ be the solution to the above dual (or concave) optimization problem. The solution satisfies the relationship (4), so that solving the dual problem for $\widehat{\mathbf{w}}$ is equivalent to finding the log-scale residuals. Hence, indirectly, via (4), we obtain the model fitted values $\mathbf{g}_{\widehat{\mathcal{Q}}}$. A challenging step is that one must solve for the parameter estimates for the model from these fitted values. We create a method that exploits the particular choice of our penalty term. Note that the constraints are linear in the parameter vector $\mathbf{w} \in \Re_+^d$, and that the number of free parameters equals d, while the number of constraints equals the cardinality of $\boldsymbol{\Theta}_m$. The dual optimization is with respect to $\mathbf{w}$ whose dimension equals d. This is especially advantageous with large data sets containing, say, thousands of observations (see Section 7.1). In Appendix A.1, we establish the relationship between the primal and dual problems at the solution.

3.1 Statistical Interpretation of the Dual Problem

The formulation in (7) and (8) is strikingly similar to the one that arises in empirical likelihood framework (Owen, 2001) in which the function $l(\mathbf{w})$ is maximized over a similar set of linear constraints. The empirical likelihood problem also has a dual problem, although it does not appear to be computationally useful.

There is a natural interpretation of the dual problem that is analogous to the linear model framework. In an application of the least squares problem, one finds the fitted values $\widehat{\mathbf{y}}$ directly by projecting the data $\mathbf{y} \in \mathcal{Y}$ onto the model space $\mathfrak{X}$ (i.e., $\mathbb{P}_{\mathfrak{X}}\,\mathbf{y} = \widehat{\mathbf{y}}$) or solves for the residual $\mathbf{e}$ by projecting $\mathbf{y}$ onto the orthogonal complement of the model space (i.e., $\mathbb{P}_{\mathfrak{X}^\perp}\,\mathbf{y} = \mathbf{e}$, where $\perp$ denotes the orthogonal projection). In turn, we solve for the fitted values using $\widehat{\mathbf{y}} = (\mathbf{y} - \mathbf{e})$. The primal and dual problems have the same relationship as the projection and complementary projection of linear models. The parallel with the linear model framework holds if we let the data $\mathbf{y}_i$ equal $\log(n_i/n)$, the fitted model $\widehat{\mathbf{y}}_i$ equal $\log\{\mathrm{g}_{\widehat{Q}}(\mathbf{y}_i)\}$ and the log-scale residual e_i equal $\log(\widehat{w}_i)$. This approach again falls very much into the spirit of the empirical likelihood, where (n_i/n) is the NPMLE of the probability of observing $\mathbf{y}_i \in \mathcal{Y}$.

3.2 The Penalized Dual Method: Theory

The goal in this section is to turn the constrained dual optimization problem defined in (7) and (8) into an unconstrained one using a "penalty function". This is referred to as the *Penalized Dual method*. Our method is in the spirit of the log-barrier method (Renegar, 2001) for convex programming; however it differs in two important respects as will be shown.

The Penalized Dual method maximizes

$$\mathcal{H}_\gamma(\mathbf{w}) = \sum_{i=1}^{d} \left(\frac{n_i}{n}\right) \log(w_i) - \mathcal{P}(\mathbf{w}, \gamma) \tag{9}$$

over $\mathbf{w} \in \Re^d_+$, where γ is a tuning parameter and $\mathcal{P}(\mathbf{w}, \gamma)$ is a *penalty function* that ensures that the *Penalized Dual solution* does not violate the constraints; the dual solution always stays in the interior of the constraint set. One choice for the penalty function is

$$\mathcal{P}(\mathbf{w}, \gamma) = \frac{1}{\gamma} \sum_{j=1}^{m} \left\{ p_{\boldsymbol{\theta}_j}(\mathbf{w}) \right\}^{\gamma} \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m \text{ and } \gamma \in \Re_+, \tag{10}$$

where the *penalty parameter* γ is some large power and the *constraint function* is defined as

$$p_{\boldsymbol{\theta}_j}(\mathbf{w}) := \sum_{i=1}^{d} w_i\, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) > 0. \tag{11}$$

That is, the dual problem constraints have the form $p_{\boldsymbol{\theta}_j}(\mathbf{w}) \leq 1$. We first show that by increasing γ, $\mathcal{P}(\mathbf{w}, \gamma)$ will eventually create an infinite penalty on any $\mathbf{w} \in \Re^d_+$ that violates the constraints and advances the solution towards the dual problem solution.

The proofs for our technical results are derived in the Appendix.

Proposition 1 For a given $\boldsymbol{\theta} \in \Omega$ and $\mathbf{w} \in \Re_+^d$, the term in the summand of the penalty function $\mathcal{P}(\mathbf{w}, \gamma)$ satisfies:

$$\frac{\{p_{\boldsymbol{\theta}}(\mathbf{w})\}^{\gamma}}{\gamma} \longrightarrow \begin{cases} \infty & \text{if} \quad p_{\boldsymbol{\theta}}(\mathbf{w}) > 1, \\ 0 & \text{if} \quad 0 \leq p_{\boldsymbol{\theta}}(\mathbf{w}) \leq 1 \end{cases}$$

as $\gamma \to \infty$. When $p_{\boldsymbol{\theta}}(\mathbf{w}) > 1$, the penalty function is increasing in γ for $\gamma > \{\log p_{\boldsymbol{\theta}}(\mathbf{w})\}^{-1}$. If $p_{\boldsymbol{\theta}}(\mathbf{w}) < 1$, the penalty function is decreasing in γ for all $\gamma \in \Re_+$.

Two main elegant features of our penalty function are the following: (1) We can directly construct an estimator for π parameters from the penalized dual solution. (2) It is simple to calculate the gradient function to assess the algorithmic convergence using the relation (20), defined in Section 3.4.

It is common in the optimization literature to employ a "barrier function" to build the penalty. For example, the *log-barrier function* defined as

$$\mathcal{P}_{\star}(\mathbf{w}, \gamma) := -\gamma \sum_{j=1}^{m} \log\{1 - p_{\boldsymbol{\theta}_j}(\mathbf{w})\} \quad \text{for} \quad j = 1, \ldots, m$$

approaches $-\infty$ as $\mathbf{w} \in \Re_+^d$ approaches the boundary of the feasible set from the interior (Roos et al., 1997; Renegar, 2001). The effect of the penalty can be diminished by making γ close to 0. Our focus here is on a soft penalty of the form (10) which is well behaved outside the feasible set; however, as will be shown, it does force the solution into the interior.

The penalized problem is unconstrained; therefore, we can find the "Penalized Dual optimal estimator" denoted $\widehat{\mathbf{w}}_{\gamma} = (\widehat{w}_{1,\gamma}, \ldots, \widehat{w}_{d,\gamma})^T$, given by (A.6) in Appendix A.2, by solving

$$\frac{\partial}{\partial w_i} \mathcal{H}_{\gamma}(\mathbf{w}) = \frac{n_i}{n} \frac{1}{w_i} - \sum_{j=1}^{m} \left\{ p_{\boldsymbol{\theta}_j}(\mathbf{w}) \right\}^{(\gamma - 1)} f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) = 0 \quad \text{for} \quad i = 1, \ldots, d. \tag{12}$$

For $\gamma = 1$, there exists an explicit solution to the above equation as

$$\widehat{w}_i \Big|_{\gamma=1} = \frac{n_i}{n} \left\{ \sum_{j=1}^{m} f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) \right\}^{-1} \quad \text{for} \quad i = 1, \ldots, d. \tag{13}$$

This is an initial interior point solution for the algorithm. On the other hand, the conventional log-barrier methods do not automatically produce a starting value for the "barrier parameter".

3.3 Existence of Parameter Estimators

We consider the following method to solve for the π parameters from the dual problem solution by exploiting the penalized structure.

Recovering the Primal Estimators: Using (4), the model fitted values $\mathbf{g}_{\widehat{\mathcal{Q}}}$ are found from the penalized dual solution $\widehat{\mathbf{w}}$. However, such a solution does not immediately provide an estimator for $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ and the technique for obtaining it is derived next.

1. Restrict attention to $\boldsymbol{\theta}_j \, (j = 1, \dots, m)$ in $\boldsymbol{\Theta}_m \subset \Omega$, the support set of $\mathcal{Q} \in \mathcal{G}$, for which the constraints are tight to ensure $\sum_{i=1}^{d} \widehat{w}_{i,\gamma} \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) = 1$.
2. Solve for $\boldsymbol{\pi}$ using the linear equations $\sum_{j=1}^{m} \widehat{\pi}_j \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) = (n_i/n)/\widehat{w}_{i,\gamma}$ for each $i = 1, \dots, d$.

The penalized dual residuals are used to obtain a natural estimator for the mixture or primal problem, denoted by $\widehat{\boldsymbol{\pi}}^{\star}_{\gamma} = \left(\widehat{\pi}^{\star}_{1,\gamma}, \dots, \widehat{\pi}^{\star}_{m,\gamma}\right)^T$. The statistic, which is referred to as the *Penalized Dual estimator* is

$$\widehat{\pi}^{\star}_{j,\gamma} = \left\{ p_{\boldsymbol{\theta}_j} \left(\widehat{\mathbf{w}}_{\gamma} \right) \right\}^{\gamma} \quad \text{for} \quad j = 1, \dots, m, \tag{14}$$

where

$$p_{\boldsymbol{\theta}_j} \left(\widehat{\mathbf{w}}_{\gamma} \right) = \sum_{i=1}^{d} \widehat{w}_{i,\gamma} \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i). \tag{15}$$

In Appendix A.2, it is shown that the estimator $\widehat{w}_{i,\gamma}$, derived in (A.6), can be approximated in terms of $\{p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_{\gamma})\}^{(\gamma-1)}$. However, these latter quantities with the power $(\gamma - 1)$ do not sum to one, and hence are turned into a *candidate estimator* via normalization:

$$\widehat{\pi}^{\dagger}_{j,\gamma} = \frac{\left\{ p_{\boldsymbol{\theta}_j} \left(\widehat{\mathbf{w}}_{\gamma} \right) \right\}^{(\gamma-1)}}{\sum_{k=1}^{m} \left\{ p_{\boldsymbol{\theta}_k} \left(\widehat{\mathbf{w}}_{\gamma} \right) \right\}^{(\gamma-1)}} \quad \text{for} \quad j = 1, \dots, m. \tag{16}$$

This candidate estimator is used to obtain $\widehat{\boldsymbol{\pi}}^{\star}_{\gamma}$ with its elements having the power γ using the following theorem.

Theorem 2 (a) For a given $\gamma \in \Re_+$, the Penalized Dual estimator

$$\widehat{\boldsymbol{\pi}}^{\star}_{\gamma} = \left[\left\{ p_{\boldsymbol{\theta}_1} \left(\widehat{\mathbf{w}}_{\gamma} \right) \right\}^{\gamma}, \dots, \left\{ p_{\boldsymbol{\theta}_m} \left(\widehat{\mathbf{w}}_{\gamma} \right) \right\}^{\gamma} \right]^T$$

is one EM-step from the candidate estimator $\widehat{\boldsymbol{\pi}}_\gamma^\dagger$; consequently, $\widehat{\boldsymbol{\pi}}_\gamma^\star$ yields a higher likelihood value. (b) The estimators are in the unit simplex $\boldsymbol{\Pi}^\star = \{\widehat{\pi}_{j,\gamma}^\star \in \Re^m : \widehat{\pi}_{j,\gamma}^\star \in [0,1], \sum_{j=1}^m \widehat{\pi}_{j,\gamma}^\star = 1\}$. (c) The Penalized Dual solution $\widehat{\mathbf{w}}_\gamma$ satisfies $p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma) \leq 1$ $(j = 1, \ldots, m)$ and hence the estimator $\widehat{\boldsymbol{\pi}}_\gamma^\star$ remains in the feasible region defined by (8).

All the proofs are relegated to Appendix A.4.

The estimator $\widehat{\boldsymbol{\pi}}^\star$ provides a direct way to obtain the primal estimator $\widehat{\boldsymbol{\pi}}$ from our penalized dual solution, avoiding the problems of selection and inversion.

3.4 Properties of the Penalized Dual Estimators

In this section, we derive several statistical properties of the estimators. First, we establish that $\widehat{\boldsymbol{\pi}}_\gamma^\star$ converges to the MLE $\widehat{\boldsymbol{\pi}}$ as the penalty parameter γ increases (Theorem 3 below). Along the way, we establish several important properties of the primal-gradient function that are necessary for solving the primal-dual problem.

Although $\widehat{\boldsymbol{\pi}}_\gamma^\star$ represents an EM improvement over $\widehat{\boldsymbol{\pi}}_\gamma^\dagger$, the candidate estimator, it is easier to establish optimization results for the latter. The following theorem shows that for a sufficiently large penalty, the Penalized Dual estimator will be close to the primal estimator. Let $\widehat{\mathcal{Q}}_\gamma^\dagger$ be the mixing distribution at the $\widehat{\boldsymbol{\pi}}_\gamma^\dagger$ solution.

Theorem 3 As $\gamma \to \infty$, $\mathbf{g}_{\widehat{\mathcal{Q}}_\gamma^\dagger} \to \mathbf{g}_{\widehat{\mathcal{Q}}}$. Consequently, the candidate estimator $\widehat{\boldsymbol{\pi}}_\gamma^\dagger$ converges to the MLE $\widehat{\boldsymbol{\pi}}$, whenever the latter is unique.

Our goal is to obtain the mixture estimation problem from the penalized dual one using $\widehat{\boldsymbol{\pi}}_\gamma^\star$. Therefore, it is important to determine directly from the dual problem how accurate is the estimator $\widehat{\boldsymbol{\pi}}_\gamma^\star$. We derive the gradient function corresponding to the mixing distribution $\widehat{\mathcal{Q}}_\gamma^\dagger$ to accomplish this. It is easier to calculate this for $\widehat{\boldsymbol{\pi}}^\dagger$ knowing that $\widehat{\boldsymbol{\pi}}_\gamma^\star$ can only be better. From the primal-gradient function in (6), the gradient function for the estimator $\widehat{\mathcal{Q}}_\gamma^\dagger$, becomes

$$\mathcal{D}_{\widehat{\mathcal{Q}}_\gamma^\dagger}(\boldsymbol{\theta}) = \sum_{i=1}^d n_i \left\{ \frac{f_{\boldsymbol{\theta}}(\mathbf{y}_i)}{\sum_{k=1}^m \widehat{\pi}_{k,\gamma}^\dagger f_{\boldsymbol{\theta}_k}(\mathbf{y}_i)} - 1 \right\} \quad \text{for} \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}_m. \tag{17}$$

Theorem 4 The primal-gradient function at the candidate estimator $\widehat{\boldsymbol{\pi}}^{\dagger}$ can be written as

$$\mathcal{D}_{\widehat{\mathcal{Q}}_{\gamma}^{\dagger}}(\boldsymbol{\theta}_j) = \frac{p_{\boldsymbol{\theta}_j}\left(\widehat{\mathbf{w}}_{\gamma}\right)}{\wp_{\gamma}} - 1 \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m \, (j = 1, \ldots, m), \tag{18}$$

where

$$\wp_{\gamma}^{-1} = \sum_{k=1}^{m} \left\{ p_{\boldsymbol{\theta}_k}(\widehat{\mathbf{w}}_{\gamma}) \right\}^{(\gamma-1)}. \tag{19}$$

Theorem 4 expresses the gradient function in terms of the dual solution and leads to a simpler device for checking the accuracy of the estimators.

Corollary 5 At the candidate estimator $\widehat{\boldsymbol{\pi}}^{\dagger}$, the primal-gradient function satisfies

$$\mathcal{D}_{\widehat{\mathcal{Q}}_{\gamma}^{\dagger}}(\boldsymbol{\theta}_j) \leq \wp_{\gamma} - 1 \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m \, (j = 1, \ldots, m), \tag{20}$$

where the term on the right-hand-side does not depend on j.

We have established that one can refine the NPMLE of $\mathcal{Q}$ to the required accuracy by increasing m and γ appropriately.

4 The Structure of the Penalized Dual Algorithm

In this section, we first investigate the structure of the penalized dual problem viewed as a function of $\mathbf{w}$ and γ and next present a strategy for their joint estimation. Next, we present the Penalized Dual algorithm to effectively search over the discretized (but large) parameter space $\boldsymbol{\Theta}_m \subset \Omega$. Lastly, convergence properties of the algorithms are derived.

We let $\mathbf{z} = \log(\mathbf{w})$ to eliminate the constraint $\mathbf{w} \in \Re_{+}^{d}$.

Theorem 6 *(a)* The function

$$\mathcal{K}(\mathbf{z}, \gamma) = \sum_{i=1}^{d} \left(\frac{n_i}{n}\right) z_i - \frac{1}{\gamma} \sum_{j=1}^{m} \left\{ p_{\boldsymbol{\theta}_j}(\mathbf{z}) \right\}^{\gamma} \quad \text{for} \quad \mathbf{z} \in \Re \quad \text{and} \quad \gamma \in \Re_{+} \tag{21}$$

is strictly concave in $(\mathbf{z}, \gamma)$, where $p_{\boldsymbol{\theta}_j}(\mathbf{z}) = \sum_i \exp(z_i)\, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)$. For any $\mathbf{z} \in \Re$ in the feasible region defined by (8), the function $\mathcal{K}(\mathbf{z}, \gamma)$ is strictly increasing as a function of γ. (b) The function $\mathcal{K}(\mathbf{z}, \gamma)$ is bounded above and achieves its maximum at $\mathbf{z} = \widehat{\mathbf{z}}$ and $\gamma = \infty$.

4.1 Automatic Selection of γ

A fundamental aspect of our algorithm is that we can maximize $\mathcal{K}(\mathbf{z}, \gamma)$ simultaneously with respect to $\mathbf{z} \in \Re$ and $\gamma \in \Re_+$; different from the approach employed in the conventional log-barrier methods in which this was not possible. Therefore, we can select the penalty parameter γ automatically. From Theorem 6, the global maximum over $\mathbf{z}$ and γ is attained when $\mathbf{z} = \widehat{\mathbf{z}}$ and $\gamma \to \infty$.

Remark 1: It may seem paradoxical to treat γ as an unknown parameter even though it has an optimum value of ∞. When γ is large, $\mathcal{K}(\mathbf{z}, \gamma)$ has very severe curvature at the constraint boundary. This limits the range of effectiveness of quadratic approximation methods. Therefore, one should start with a small value for γ and increase it as the algorithm progresses through the parameter space. This could possibly be achieved in some other systematic fashion; however, our empirical investigations suggest that systematic methods were not as efficient as our approach. A possible explanation could be that our strategy takes the curvature of the function $\mathcal{K}(\mathbf{z}, \gamma)$ into account, in providing the relevant information for determining the increments for γ.

4.2 Searching Effectively Over the Discretized Parameter Space

An algorithm for efficiently searching over the large discretized parameter space $\mathbf{\Theta}_m \subset \Omega$ (required for Step 1 of Algorithm 1 described in Section 2.1) is derived next. In effect, the following algorithm is used for fitting the fixed support mixture model.

Algorithm 2 (The Penalized Dual Algorithm)

1. Consider $\gamma = 1$ and its corresponding explicit solution $\widehat{\mathbf{z}}^{(1)}$ given in (13) as the starting solution for the algorithm.
2. Maximize the concave function $\mathcal{K}(\mathbf{z}, \gamma)$ simultaneously with respect to $\mathbf{z} \in \Re$ and $\gamma \in \Re_+$ using a modified Newton-Raphson algorithm [constrained the step size to ensure monotonicity in $\mathcal{K}(\mathbf{z}, \gamma)$] until the following convergence criterion is satisfied; namely, the L_2-norm of the change in the value of $\mathcal{K}(\mathbf{z}, \gamma)$ is less than 10^{-6}.

3. Fix γ at $\gamma^{(k)}$ obtained in Step 2 and find $\widehat{\mathbf{z}}_{\gamma^{(k)}} = \arg\max_{\mathbf{z}\in\Re} \mathcal{K}\left(\mathbf{z}, \gamma^{(k)}\right)$ using the modified Newton-Raphson algorithm. The algorithm is considered to have converged to the maximum at step t, when the inequality based on the primal-gradient function

$$\Psi\left(\mathcal{Q}^{(t)}\right) \leq 0.005 \tag{22}$$

is satisfied since it guarantees convergence to a similar accuracy in the loglikelihood.

As described in Section 2, the supremum of the gradient function $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m} \mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta})$ provides an assessment of the progression to the maximum and hence the criterion in (22) has a solid theoretical justification.

Remark 2: The Step 3 of the algorithm is necessary since after Step 2, the Primal Dual estimator $\widehat{\boldsymbol{\pi}}^{\star}_{\gamma}$ obtained via (14) are often not sufficiently close to the primal estimator $\widehat{\boldsymbol{\pi}}$. This is because the algorithm does not necessarily satisfy the condition $\{\partial\mathcal{K}(\mathbf{z},\gamma)/\partial\mathbf{z}\} = 0$ with sufficient accuracy. In our applications, however, the primal-gradient inequality (22) was always achieved at the tolerance of 0.005; in fact, often reached significantly greater accuracy in the Penalized Dual estimators.

The Penalized Dual Algorithm with Inactive Constraints: In Algorithm 2, if an estimated mixture probability $\widehat{\pi}_j$ $(j = 1, \ldots, m)$ is zero, then the corresponding constraint in the dual problem is inactive. We can dynamically update the active constraints by removing the inactive ones while adding new ones, whenever the support set violated the gradient inequality. From the Penalized Dual estimator in (14), it follows that if $\widehat{\pi}_j \to 0$, then $\{p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma)\}^\gamma \to 0$ which occurs when $p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma) \to 0$ or $\gamma \to \infty$. In the former, one can essentially remove the corresponding density $f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)$. It will be shown in Section 7 that the above algorithm, denoted by $\mathrm{PD}^{\mathrm{IC}}$, produced a further reduction in computational time.

As a consequence of the concavity of $\mathcal{K}(\mathbf{z},\gamma)$ established in Theorem 6, the Hessian $\mathbf{H}$ for $\mathcal{K}(\mathbf{z},\gamma)$ (derived in equation (A.7) in Appendix A.3) is always non-singular and the sequence obtained from the Penalized Dual algorithm (i.e., Algorithm 2) are well defined. Even for large-scale problems such as the yeast microarray data considered in Section 7.3 in which $\mathbf{H}$ is of dimension 697, our modified Newton-Raphson algorithm was stable and efficient.

Theoretically, owing to Theorem 6, the Step 2 of the Penalized Dual algorithm produces a sequence $\left\{\mathbf{z}^{(k)}, \gamma^{(k)}\right\}_{k\geq 1}$ such that the sequence of functions $\left\{\mathcal{K}(\mathbf{z}^{(k)}, \gamma^{(k)})\right\}_{k\geq 1} \to \mathcal{K}(\widehat{\mathbf{z}}, \infty)$ as $k \to \infty$. This effectively implies that the sequence $\left\{\mathbf{z}^{(k)}\right\}_{k\geq 1} \to \widehat{\mathbf{z}}$ and the sequence $\left\{\gamma^{(k)}\right\}_{k\geq 1} \to \infty$ as $k \to \infty$. However, in practice, convergence of γ is slow; therefore, we terminate the modified Newton-Raphson algorithm in Step 2 when $\gamma^{(k)}$ is sufficiently large and maximize $\mathcal{K}\left(\mathbf{z}, \gamma^{(k)}\right)$ over $\mathbf{z}$ for a fixed $\gamma^{(k)}$.

In our experience, a direct maximization of the mixture loglikelihood $l(\boldsymbol{\pi})$ over Π using a modified Newton-Raphson algorithm was unstable and failed to converge to the maximum $\widehat{\boldsymbol{\pi}}$.

5 Convergence Properties of the Algorithms

In this section, we establish the convergence properties, including the rate of convergence, of the algorithms.

First, we consider the algorithm for fitting the continuous support mixture model; i.e., an algorithm employed in Step 2 of Algorithm 1. We prove that the sequence of estimates $\{\boldsymbol{\beta}^{(s)}\}_{s\geq 1}$ obtained from the Step 2 of Algorithm 1 converges to an MLE of $\boldsymbol{\beta} \in \Omega$, namely $\widehat{\boldsymbol{\beta}}$, for a given data $\mathbf{y} \in \mathcal{Y}$ as s increases. For instance, in the multivariate normal mixture framework, $\boldsymbol{\beta}$ becomes $(\mathcal{Q}, \boldsymbol{\Sigma})$. Assume that the sequence of estimates $\{\boldsymbol{\beta}^{(s)}\}_{s\geq 1}$ monotonically increases the loglikelihood $l(\boldsymbol{\beta})$. An algorithm is said to converge if $\boldsymbol{\beta}^{\star} = \lim_s \boldsymbol{\beta}^{(s)}$ exists, for a parameter vector $\boldsymbol{\beta} \in \Omega$.

Wu (1983) established that monotonicity of $l(\cdot)$ does not imply the convergence of the sequence to a stationary point; however, if the sequence $\{l(\boldsymbol{\beta}^{(s)})\}_{s\geq 1}$ is bounded above, then it does converge monotonically to a stationary point of $l(\boldsymbol{\beta})$. The convergence of $\boldsymbol{\beta}^{(s)}$ to $\widehat{\boldsymbol{\beta}}$ implies the convergence of $l(\boldsymbol{\beta}^{(s)})$ to $l(\widehat{\boldsymbol{\beta}})$ according to the Theorem 5, under the regularity conditions, derived by Wu (1983).

Owing to Theorem 6, the Algorithm 2 (or Step 1 of Algorithm 1) produces a sequence of estimates $\{\boldsymbol{\pi}^{(t)}\}_{t\geq 1}$ that is guaranteed to converge to the unique MLE $\widehat{\boldsymbol{\pi}}$. Combined this result with Theorem 5 in Wu (1983) establishes the convergence of Algorithm 1 to $\widehat{\boldsymbol{\beta}}$.

5.1 Convergence Criteria

For the applications and simulation experiment, we used the convergence criterion based on the gradient function for the Penalized Dual (PD) and discrete EM (i.e., for fitting the fixed support mixture model) algorithms. That is, the algorithm has converged to the MLE $\widehat{\boldsymbol{\pi}}$ if the criterion in (22) is satisfied. For the rest of the article, we denote the discrete EM by D-EM algorithm.

The D-EM algorithm is a sublinearly convergent algorithm (Pilla and Lindsay, 2001); therefore, a conventional convergence criterion based on the loglikelihood change or changes in parameters, such as

$$\xi^{(t)} = \left| l\left(\boldsymbol{\pi}^{(t)}\right) - l\left(\boldsymbol{\pi}^{(t-1)}\right)\right| \leq \tau \tag{23}$$

for a given tolerance τ can be very misleading in the sense that the actual distance to the final loglikelihood

$$\Lambda^{(t)} = \left| l\left(\widehat{\boldsymbol{\pi}}\right) - l\left(\boldsymbol{\pi}^{(t)}\right)\right| \tag{24}$$

can be orders of magnitude different from τ. That is, this criterion may be met even though the parameter values are far from the correct solution (Titterington et al., 1985; Pilla and Lindsay, 2001). However, such rules are widely employed and therefore we conducted an experiment to assess the two criteria on two data sets.

The most important assessment of the convergence of an ML algorithm is the value of the loglikelihood, as it provides information about the accuracy of parameter estimators on a confidence interval scale. Therefore, loglikelihood-based criterion is a useful one to employ in assessing the convergence of an algorithm in finding the MLE of the parameters (Lindsay, 1995; Pilla and Lindsay, 2001).

Simulation Experimental Design: We consider the simulated data by generating a sample of size $n = 270$ from $N_p(\mathcal{Q}, \mathbf{I})$ with $p = 3$, where $N_p(\mathcal{Q}, \mathbf{I})$ represents a measure of a p-dimensional normal random variable with mean $\mathcal{Q}$ and an identity variance-covariance matrix. The true mixing measure for $\mathcal{Q} \in \mathcal{G}$, is chosen by selecting the coordinates of $\boldsymbol{\theta}_j\,(j = 1, \ldots, m)$ from the set $\{-5, 0, 5\}$ in all possible combinations, with equal mass at each support vector. This resulted in a total of $m = 3^3$ mixture components.

Fisher Iris Data: We fit a mixture of multivariate normal distributions to Fisher iris data (Fisher, 1936). The data consists of $n = 150$ observations collected on flowers of three iris species (Setosa, Verginica and Versicolor). Each observation is a vector of $p = 4$ variables sepal length ($\mathbf{y}_1$), sepal width ($\mathbf{y}_2$), petal length ($\mathbf{y}_3$) and petal width ($\mathbf{y}_4$).

Table 1: Effect of a convergence criterion on the final loglikelihood in fitting the fixed support mixture model.

Experiment	t	$l\left(\boldsymbol{\pi}^{(t)}\right)$	$\Psi\left(\mathcal{Q}^{(t)}\right)$	$\Lambda^{(t)}$
Simulated	1067	-2313.6826	0.0830	0.0536
Fisher Iris	460	-376.9595	3.0017	0.0156

For each of the data sets, we selected the observed data matrix $\mathbf{y}$ for $\boldsymbol{\Theta}_m$ and also set $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$, the sample variance-covariance matrix. In order to assess the accuracy of the algorithms at a given step t, we found the final loglikelihood value $l(\widehat{\boldsymbol{\pi}})$ to a high degree of accuracy using the PD algorithm for a sufficiently large t. Next, we fit mixtures of multivariate normal distributions to the simulated and iris data sets via the D-EM algorithm using the convergence criterion (23) with $\tau = 0.0001$. The $\Lambda^{(t)}$ values, presented in Table 1, demonstrate that the convergence criterion (23) would result in substantially less than four decimal accuracy for the Fisher iris data. On the other hand, the criterion based on $\Psi(\mathcal{Q}^{(t)})$ in (22) guarantees the final accuracy.

5.2 Empirical Assessment of Convergence Rate

We empirically assess the rate of convergence of the PD algorithm relative to the D-EM algorithm by defining $\Lambda^{(t)}$ in (24) as the *residual of the loglikelihood* at the tth step.

To be precise, for some $\boldsymbol{\pi}^{(0)} \in \boldsymbol{\Pi}$, let $\left\{\boldsymbol{\pi}^{(t)}\right\}_{t\geq 1}$ be a sequence in $\boldsymbol{\Pi}$ generated by an algorithm (such as the PD and D-EM algorithms). The algorithm can be expressed as $\boldsymbol{\pi}^{(t)} \in \mathcal{M}\left(\boldsymbol{\pi}^{(t-1)}\right)$ for $t \geq 1$, where the map $\mathcal{M} : \boldsymbol{\Pi} \to 2^{\boldsymbol{\Pi}}$ is a point-to-set mapping. If $\boldsymbol{\pi}^{(t)}$ converges to $\widehat{\boldsymbol{\pi}}$ and $\mathcal{M}(\cdot)$ is continuous, then $\widehat{\boldsymbol{\pi}}$ must satisfy $\widehat{\boldsymbol{\pi}} \in \mathcal{M}(\widehat{\boldsymbol{\pi}})$.

Definition 2 (Asymptotic Convergence Rate): Assume that $\widehat{\boldsymbol{\pi}} = \mathcal{M}(\widehat{\boldsymbol{\pi}})$ and that the sequence $\{\boldsymbol{\pi}^{(t)}\}_{t\geq 1}$ is generated by the map $\mathcal{M}$ such that $\lim_{t\to\infty} \boldsymbol{\pi}^{(t)} = \widehat{\boldsymbol{\pi}}$. Under the regularity conditions given by Wu (1983), this implies that $\lim_{t\to\infty} l\left(\boldsymbol{\pi}^{(t)}\right) = l(\widehat{\boldsymbol{\pi}})$. The *asymptotic convergence rate of the loglikelihood sequence* $\left\{l\left(\boldsymbol{\pi}^{(t)}\right)\right\}_{t\geq 1}$ at $l(\widehat{\boldsymbol{\pi}})$ generated by an algorithm is defined as

$$r := \lim_{t\to\infty} \left| l\left(\widehat{\boldsymbol{\pi}}\right) - l\left(\boldsymbol{\pi}^{(t)}\right) \right|^{\frac{1}{t}}$$

From the following lemma (Pilla and Lindsay, 2001), the smaller the r for any given loglikelihood sequence, the faster it is progressing towards the MLE.

Lemma 7 If the sequence $\{l(\boldsymbol{\pi}^{(t)})\}_{t\geq 1}$ is converging linearly, then as $t \to \infty$, the slope of the curve obtained by plotting $\log\left\{\Lambda^{(t)}\right\}$ against t converges to $\log(r)$, where r is the asymptotic rate of convergence of the loglikelihood sequence generated by an algorithm.

In order to assess the rate of convergence of the PD, relative to the D-EM, algorithm, we consider the Fisher iris data considered earlier for fitting a collection of semiparametric mixture of multivariate normal distributions $N_p(\mathcal{Q}, \delta\,\boldsymbol{\Sigma})$, where $p = 4, \boldsymbol{\Sigma} \in \mathcal{S}$ and $\delta \in \Re_+$ (details in Section 6.2). As before, we selected the observed data $\mathbf{y}$ for $\boldsymbol{\Theta}_m$ and set $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$ in fitting the PD and D-EM algorithms. Figure 1 demonstrates the behavior of the algorithms for $\delta \in \{5, 2, 1, 0.5\}$. We used logarithmic scaling of the vertical axis since a linearly convergent algorithm will become linear on this scale as $t \to \infty$. Note that for the fixed support mixture model, the D-EM algorithm is converging sublinearly whereas the PD is converging linearly to the MLE $\widehat{\boldsymbol{\pi}}$; a significant improvement in convergence rate. In fact, Pilla and Lindsay (2001) observed a similar behavior of sublinear convergence of the D-EM algorithm for a class of univariate finite mixture problems.

6 Semiparametric Mixtures of Multivariate Normal Distributions

The methodology developed in this article is applicable to a wide range of problems, including multivariate t mixtures. However, the particular interest here is in difficult problems

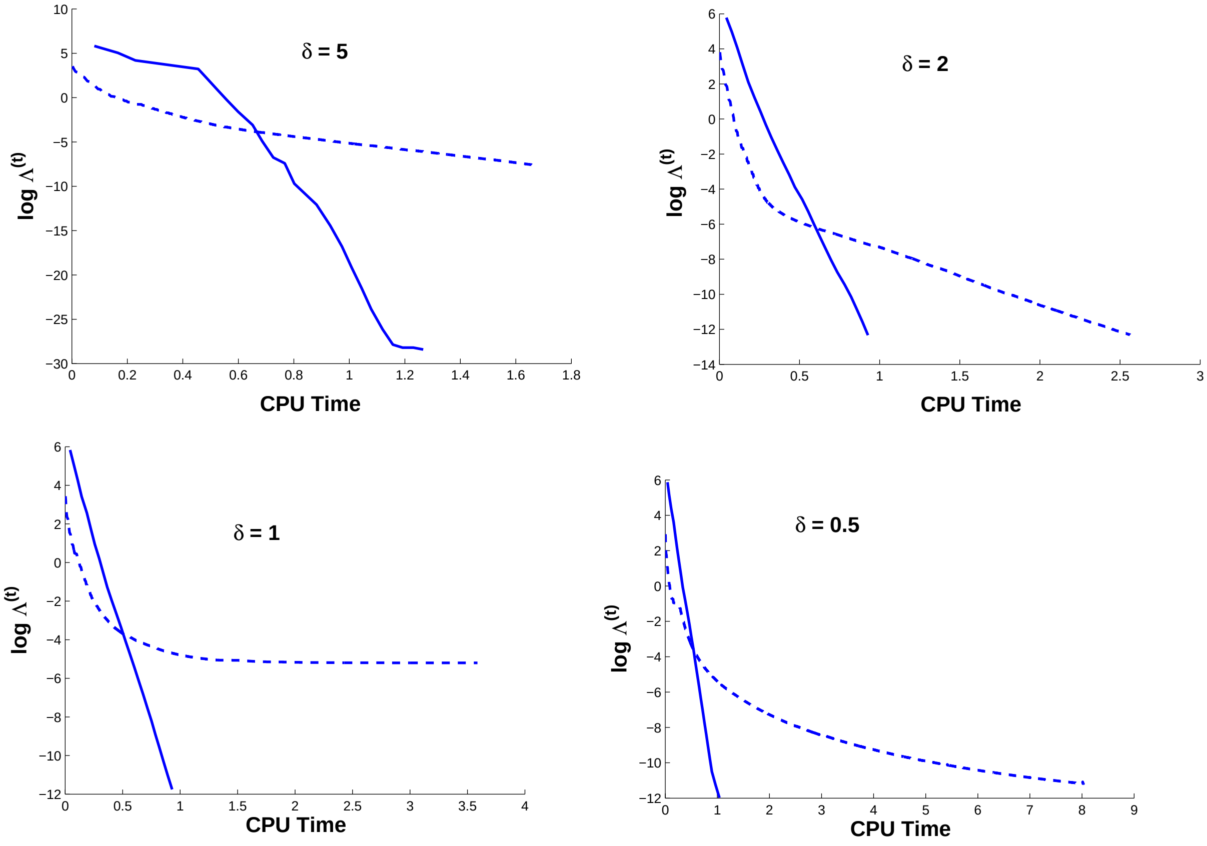

Figure 1: Plot of the log residual of the loglikelihood, $\log\{\Lambda^{(t)}\}$ against CPU time for the Fisher iris data with four variables demonstrating the sublinear convergence of the discrete EM (dashed line) and the linear convergence of the Penalized Dual (solid line) algorithms.

with multivariate normal mixtures due to its ubiquitous applications. In semiparametric mixture setting, the mixing distribution $\mathcal{Q}$ is modeled nonparametrically in the presence of an unknown $\boldsymbol{\Sigma} \in \mathcal{S}$, the variance-covariance matrix common to all m components.

6.1 Structural Properties

Let $\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i;\boldsymbol{\Sigma}) = \sum_{j=1}^{m} \pi_j\, f_{\boldsymbol{\mu}_j}(\mathbf{y}_i;\boldsymbol{\Sigma})$ for $\mathbf{y}_i \in \mathcal{Y}$ be a finite mixture of p-dimensional normal distributions, where $\boldsymbol{\mu}_j \in \Re^p \subset \Omega$ is the mean vector of the jth component density $f_{\boldsymbol{\mu}_j}(\mathbf{y}_i;\boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \in \mathcal{S}$ is common to all m components. Note that in the continuous case $d = n$. The corresponding loglikelihood is expressed as

$$l(\mathcal{Q};\mathbf{\Sigma}) = \sum_{i=1}^{n} \log\{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i;\mathbf{\Sigma})\}. \tag{25}$$

In the univariate case, Charnigo and Pilla (2005) establish that for a general family of mixture models with a structural parameter β (e.g., σ^2 in the normal case), the likelihood framework breaks down when joint estimation of $m, \mathcal{Q}$ and β is attempted: at best the joint estimator of $m, \mathcal{Q}$ and β is degenerate, and at worst it does not even exist. The ML fails in this setting since taking finite samples from continuous probability distributions yields discrete data sets. When models that closely mimic discrete probability distributions are available, as they are when there are no restrictions on $\mathcal{Q}$ and β, the likelihood will favor such models. The NPMLE results of Lindsay (1995, Section 2.6) cannot be applied if $\mathbf{\Sigma}$ is unknown; however, the following result holds.

We define the gradient function for the multivariate normal mixture distributions as

$$\mathcal{D}_{\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}}(\boldsymbol{\mu};\mathbf{\Sigma}) := \sum_{i=1}^{n}\left\{\frac{f_{\boldsymbol{\mu}}(\mathbf{y}_i;\mathbf{\Sigma})}{\mathrm{g}_{\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}}(\mathbf{y}_i;\mathbf{\Sigma})} - 1\right\} \quad \text{for} \quad \boldsymbol{\mu}\in\Omega. \tag{26}$$

Next, we define an NPMLE of $\mathcal{Q}\in\mathcal{G}$ for a fixed $\mathbf{\Sigma}\in\mathcal{S}$ as

$$\widehat{\mathcal{Q}}_{\mathbf{\Sigma}} = \arg\max_{\mathcal{Q}\in\mathcal{G}} l(\mathcal{Q};\mathbf{\Sigma}).$$

Theorem 8 (Unique NPMLE of $\mathcal{Q}$) Assume $\mathbf{\Sigma} > 0$ is fixed.

(1) Suppose $\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}$ satisfies

$$\mathcal{D}_{\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}}(\boldsymbol{\mu};\mathbf{\Sigma}) \le 0 \quad \text{for all} \quad \boldsymbol{\mu}\in\Omega, \tag{27}$$

then $\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}$ is an NPMLE of $\mathcal{Q}\in\mathcal{G}$.

(2) Let the set $\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}$ for some $K \le n$ be the solution set

$$\left\{\boldsymbol{\mu} : \mathcal{D}_{\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}}(\boldsymbol{\mu};\mathbf{\Sigma}) = 0\right\}.$$

If the vectors

$$\mathbf{f}_{\boldsymbol{\mu}_j}(\mathbf{y};\mathbf{\Sigma}) = \left\{f_{\boldsymbol{\mu}_j}(\mathbf{y}_1;\mathbf{\Sigma}),\ldots,f_{\boldsymbol{\mu}_j}(\mathbf{y}_n;\mathbf{\Sigma})\right\}^T \text{ for } j = 1,\ldots,K$$

are linearly independent, then $\widehat{\mathcal{Q}}_{\mathbf{\Sigma}}$ is the unique NPMLE of $\mathcal{Q}\in\mathcal{G}$.

For the multivariate normal mixture model with a common $\boldsymbol{\Sigma}$, we restrict attention to finite discrete latent distributions $\mathcal{Q}$, then the pair $(\mathcal{Q}, \boldsymbol{\Sigma})$ is identifiable (Lindsay, 1995). For a general family of univariate mixtures, Charnigo and Pilla (2005) establish that joint estimation of $m, \mathcal{Q}$ and β is a *well-defined* problem if $\mathcal{Q}$ is finitely supported. If $\mathcal{Q}$ is not finitely supported, then $\mathrm{g}_{\mathcal{Q}}(\mathbf{y}; \boldsymbol{\Sigma})$ need not determine $\mathcal{Q}$ and β uniquely. Hence, an ML approach to the joint estimation of $m, \mathcal{Q}$ and $\boldsymbol{\Sigma}$ fails. However, since we fix m and consider $\mathcal{Q}$ to be finitely supported, joint estimation of $\mathcal{Q}$ and $\boldsymbol{\Sigma}$ is feasible. Therefore, we can apply Algorithm 1 described in Section 2.1 to jointly estimate $\mathcal{Q}$ and $\boldsymbol{\Sigma}$.

The following theorem establishes that joint identifiability of $(\mathcal{Q}, \boldsymbol{\Sigma})$ fails if $\mathcal{Q}$ is not finitely supported. The proof follows from the univariate nesting structure result, under mild regularity conditions, given by Charnigo and Pilla (2005).

Theorem 9 (Multivariate Mixture Nesting Structure) The class of multivariate normal mixture distributions possesses the nesting structure. That is, for any $\boldsymbol{\Sigma}^{\dagger} \succ \boldsymbol{\Sigma}$, in the sense of Löwner ordering,

$$\{N_p(\mathcal{Q}, \boldsymbol{\Sigma}^{\dagger})\colon \mathcal{Q} \in \mathcal{G}\} \subseteq \{N_p(\mathcal{Q}, \boldsymbol{\Sigma})\colon \mathcal{Q} \in \mathcal{G}\},$$

where $N_p(\mathcal{Q}, \boldsymbol{\Sigma})$ represents a measure of a p-dimensional normal random variable with mean $\mathcal{Q}$ and a variance-covariance matrix $\boldsymbol{\Sigma} \in \mathcal{S}$.

6.2 Role of the Sieve Parameter in Building Semiparametric Mixture Models

We investigate building the sieve of models $N_p(\mathbf{F}, \, \delta \, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathcal{S}$ and $\delta \in \Re_+$ is a *sieve parameter* (similar to the smoothing parameter employed in density estimation). The sieve parameter controls the dimensionality of a mixture model as will be demonstrated later. We derive theory for building a collection of semiparametric mixture models, including the multivariate case.

In order to create a general family of mixture models, we consider the class $\{N_p(\mathcal{Q}, \, \delta \, \boldsymbol{\Sigma}) : \mathcal{Q} \in \mathcal{G}, \boldsymbol{\Sigma} \in \mathcal{S}, \delta \in \Re_+\}$. For $\delta_1 > \delta_0$, Theorem 9 implies that

$$\{N_p(\mathcal{Q}, \, \delta_1 \, \boldsymbol{\Sigma})\colon \mathcal{Q} \in \mathcal{G}\} \subseteq \{N_p(\mathcal{Q}, \, \delta_0 \, \boldsymbol{\Sigma})\colon \mathcal{Q} \in \mathcal{G}\};$$

hence, the collection of models becomes richer as $\delta \to 0$ (see Figures 2 and 3). Moreover, every p-dimensional distribution $\mathbf{F}$ can be obtained as the weak limit of $N_p(\mathbf{F},\, \delta\,\mathbf{\Sigma})$ as $\delta \to 0$. Therefore, we can approximate any distribution by choosing δ small. As a consequence, the principle of maximum likelihood cannot be applied to select δ in the model $N_p(\mathcal{Q},\, \delta\,\mathbf{\Sigma})$ since the likelihood becomes unbounded as $\delta \to 0$. Charnigo and Pilla (2005) develop theory for the univariate mixtures and demonstrate the effect of small δ for a general family of univariate mixtures which extends to the multivariate case considered here.

We create a strategy for building a collection of models $\{(\mathcal{Q}_\delta,\, \delta\,\mathbf{\Sigma})\colon \mathcal{Q} \in \mathcal{G}, \delta \in \Re_+\}$ using the Penalized Dual algorithm. As $\delta \to 0$, the NPMLE $\widehat{\mathcal{Q}}_\delta$ converges in distribution to $n^{-1}\sum_i \vartheta(\mathbf{y}_i)$, where $\vartheta(\mathbf{y})$ is a discrete measure concentrated at $\mathbf{y}$.

To demonstrate the effect of δ on the mixture complexity, we create a collection of models for both the univariate and multivariate data. In the univariate case, we simply have a σ parameter. The univariate application considers the galaxy data set [Table 1 of Roeder (1990)] of 82 observations of relative velocities for galaxies from six well separated conic sections of the Corona Borealis region. Scientific interest lies in identifying substructures in clusters of galaxies. Multimodality is evidence of voids and superclusters in the far universe. Roeder (1990) obtained $\widehat{\sigma} = 0.95$ using least squares cross validation. We set $\boldsymbol{\mu} \in \mathbf{\Theta}_m = \{9, \ldots, 35\}$ with a grid size of 0.02 for building a collection of semiparametric mixture models using Algorithm 2. The plot of $\log(\sigma)$ against the support set $\boldsymbol{\mu}$ corresponding to the estimate $\widehat{\boldsymbol{\pi}}$ obtained using the PD algorithm (namely, Algorithm 2) is shown in Figure 2. That is, at each fixed $\log(\sigma)$, the plot displays μ parameter values that have positive mixture probability. The figure demonstrates the effect of σ on the mixture complexity m.

Next, we consider the Fisher's iris data described earlier. Once again, we selected observed data $\mathbf{y}$ for $\mathbf{\Theta}_m$ and set $\widehat{\mathbf{\Sigma}} = \mathbf{S}$. We let $\delta = \{0.1, \ldots, 5\}$ for building a collection of semiparametric multivariate mixture models using Algorithm 2. Figure 3 shows the effect of δ on the mixture model complexity when only the two variables, namely the petal length and petal width are considered. The galaxy and Fisher iris data sets demonstrate that the number of components is a consequence of the choice of σ (or δ as the case may be) rather than a pre-selected parameter.

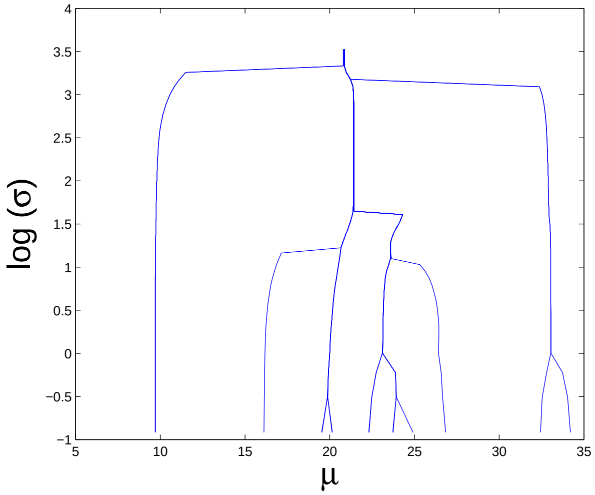

Figure 2: A Mixture Tree for the galaxy data set.

6.3 Selection of the Support Set of $\mathcal{Q}$

As described in Section 2.1, in approximating $\mathcal{G}$, the biggest challenge is in selecting a suitable $\mathbf{\Theta}_m$ while keeping computations manageable. This is addressed in this section.

In the absence of a prior knowledge of the mixture complexity, correct specification of $\mathbf{\Theta}_m \subset \Omega$, the support set of $\mathcal{Q}$, is very important for the Step 1 of Algorithm 1 (or equivalently for Algorithm 2). As expected, the final loglikelihood depends on this choice. In this section, we illustrate through the simulated data described earlier how the observed data matrix $\mathbf{y}$ provides the best choice for approximating the continuous parameter space Ω. In effect, we select $\{\boldsymbol{\theta}_1 = \mathbf{y}_1, \boldsymbol{\theta}_2 = \mathbf{y}_2, \ldots, \boldsymbol{\theta}_m = \mathbf{y}_d\}$. Note that choosing $\mathbf{y}$ for the discrete parameter space $\mathbf{\Theta}_m$ clearly covers the region of likely support vectors for the normal means and has the advantage of adapting naturally in richness to the sample size of the problem.

To assess the effectiveness of using $\mathbf{y}$ for $\mathbf{\Theta}_m$ (which is approximating the continuous parameter space Ω) we consider the simulated data described in Section 5.1. The true mixing measure for $\mathcal{Q}$ chosen for the simulation experiment is denoted by “True Support” in Table 2. The “Equi-Distant” set for $\mathbf{\Theta}_m$ was constructed on a lattice by choosing the elements in $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \theta_{j3})$ for $j = 1, \ldots, 8^3$ from the set $\{-7, -5, \ldots, 5, 7\}$ resulting in a total of $m = 8^3$ support vectors; this set also included all the true support vectors. Table 2 presents

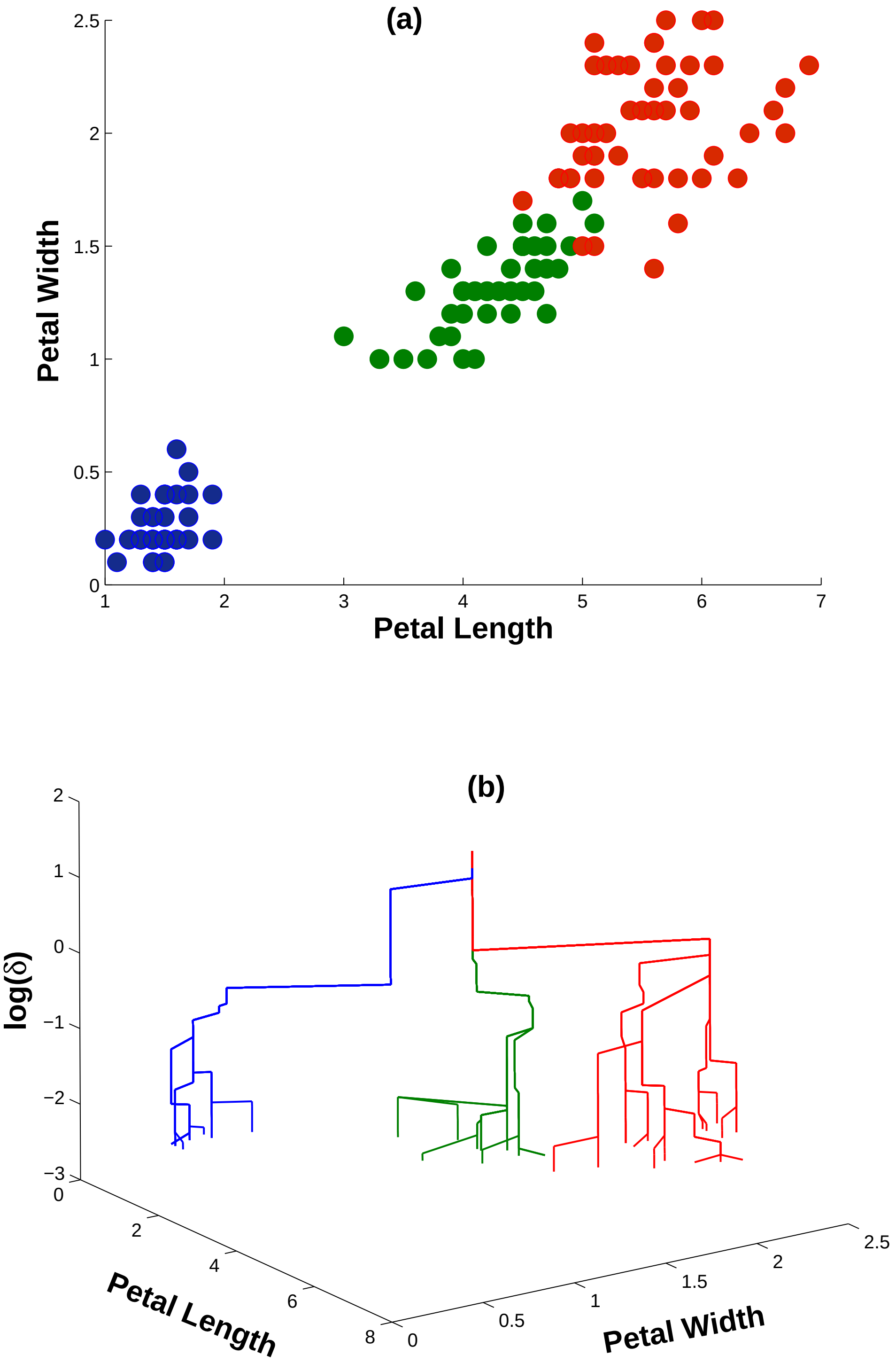

Figure 3: (a) The scatter plot of the Fisher iris data for the two variables, namely the petal length and petal width, showing three main groups. (b) A Mixture Tree demonstrating the effect of the sieve parameter δ on the number of components. The three main branches in the tree correspond to the three main components in (a).

results obtained using Algorithm 2 (i.e., fixed support mixture model of estimating $\boldsymbol{\pi}$ for a given $\boldsymbol{\Theta}_m$) and Step 2 of Algorithm 1 (i.e., continuous support mixture model of estimating $\mathcal{Q}$ and $\boldsymbol{\Sigma}$ for a fixed m).

In general, $\mathbf{y}$ should be an effective choice for $\boldsymbol{\Theta}_m$ given that equi-distant is still a subjective one in the absence of any knowledge about the length of the distance. From the theory presented in Section 2.1, as $m \to \infty$, $\boldsymbol{\Theta}_m \to \Omega$. However, in practice, choosing $m = n$ is effectively creating a dense set for $\boldsymbol{\Theta}_m$ and in fact approximating Ω very well.

Table 2: Effect of $\boldsymbol{\Theta}_m \subset \Omega$, the support set for $\mathcal{Q}$, on the estimated loglikelihood. We set $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$ in finding $l(\widehat{\boldsymbol{\pi}})$ using the Penalized Dual (PD) algorithm. The solution $\widehat{\boldsymbol{\pi}}$ obtained from the PD algorithm with the corresponding fixed support set and $\mathbf{S}$ are employed as parameter starting values for finding $l\left(\widehat{\mathcal{Q}}_\delta, \delta\,\widehat{\boldsymbol{\Sigma}}\right)$ for a fixed $\delta = 0.2$ using the continuous EM (C-EM) algorithm.

$\boldsymbol{\Theta}_m$	$l\left(\widehat{\boldsymbol{\pi}}\right)$	$l\left(\widehat{\mathcal{Q}}_\delta, \delta\,\widehat{\boldsymbol{\Sigma}}\right)$
True Support	-2181.9	-1936.9
Equi-Distant	-2182.8	-1901.7
Observed Data	-2178.6	-1876.0

7 Applications and Simulation Experiment

The applications in this section are used to investigate the roles of many overlapping components which create an ideal situation for solving the large-scale practical problems.

We assess the performance of the algorithms in finding the NPMLE of $\mathcal{Q}$ and for fitting the collection of semiparametric mixture models with applications to several data sets. The data sets, the parameter estimates and the Matlab software for fitting mixtures are available from the first author. For the Step 1 of Algorithm 1, we set $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$; however, we estimate it in Step 2.

7.1 Mortality Data

Our first application illustrates the tremendous advantage of our method in reducing the dimension of discrete mixture problems. In these problems the magnitude of d, the number of distinct observed data points, could be much smaller than m, the cardinality of $\boldsymbol{\Theta}_m$.

We consider the data on death rates which gives the number of death notices for women aged 80 and over, from the *Times* newspaper for each day in the three-year period 1910 to 1912 (Titterington et al., 1985). For the later data sets, we chose the observed data matrix $\mathbf{y}$ as the support set $\boldsymbol{\Theta}_m$. However, for this application, we selected the support set to be $\boldsymbol{\Theta}_m = \{0, 0+\eta, \ldots, 9-\eta, 9\}$, where $\eta \in \{1, 0.5, 0.1, 0.01\}$. In effect, the mixture complexity $m \in \{10, 20, 100, 1000\}$. It is worth noting that the dimension of the dual optimization problem is d (equals 10) whereas that of the mixture problem is $(m-1)$ which grows significantly with the cardinality of the set η.

We fit a mixture of Poisson distributions to the mortality data using the PD and D-EM algorithms. Table 3 presents $\mathrm{N}^{(t)}$, the number of steps required for convergence [based on the criterion (22)] and "CPU Factor", the ratio of the CPU time required by the D-EM algorithm to that of the PD. This ratio indicates the factor by which the D-EM algorithm is accelerated. We also present the values of $l(\widehat{\boldsymbol{\pi}}), l(\widehat{\mathcal{Q}}_\eta), \Lambda^{(t)}$ and $\Psi(\mathcal{Q}^{(t)})$. Furthermore, we consider the effect of eliminating the inactive constraints in the PD algorithm, namely $\mathrm{PD}^{\mathrm{IC}}$. The table demonstrates that the PD-based algorithms advance toward the maximum more rapidly than does the D-EM algorithm with gains increasing as m (equivalently, the number of parameters to estimate) increases. Thousand-fold improvements are obtained at $\eta = 0.01$ for which the number of parameters to estimate is the largest. For comparison, we fit the same model with the Rotated EM (an accelerated version of the EM applicable only for univariate mixtures) developed by Pilla and Lindsay (2001) and obtained CPU factors for the PD, relative to the Rotated EM, as 1.4, 22, 43 and 28, respectively for $\eta \in \{1, 0.5, 0.1, 0.01\}$.

At $\eta = 0.01$, the D-EM has retained 199 support points with non zero probability at convergence (obtaining a smaller loglikelihood value) whereas the PD has retained just 26 support points and reached the MLE in a reasonable number of steps; a significant reduction

Table 3: Building a collection of finite mixture models $\left\{\widehat{\mathcal{Q}}_\eta : \eta \in \Re_+\right\}$ using Algorithm 1 for the mortality data. First step involved setting $\mathbf{\Theta}_m = \{0, 0+\eta, \ldots, 9-\eta, 9\}$ with $\eta \in \{1, 0.5, 0.1, 0.01\}$ and finding $l(\widehat{\boldsymbol{\pi}})$ using the PD-based and discrete EM (D-EM) algorithms. The estimate $\widehat{\boldsymbol{\pi}}$ obtained from the PD algorithm with the corresponding fixed support set is used as parameter starting values for finding $l\left(\widehat{\mathcal{Q}}_\eta\right)$ using the continuous EM (C-EM) algorithm.

		$l(\widehat{\boldsymbol{\pi}})$	$\Lambda^{(t)}$	$\Psi\left(\mathcal{Q}^{(t)}\right)$		**CPU**
Algorithm	η	$l\left(\widehat{\mathcal{Q}}_\eta\right)$	$\times 10^3$	$\times 10^3$	$\mathrm{N}^{(t)}$	**Factor**
PD	1	-1990.0928	0.0000	0.2577	25	5
$\mathrm{PD}^{\mathrm{IC}}$		-1990.0928	0.0000	0.2577	25	7
D-EM		-1990.0929	0.0172	4.9885	1,238	1
C-EM		-1989.9	-	-	2,179	-
PD	0.5	-1989.9941	0.0000	0.1881	26	120
$\mathrm{PD}^{\mathrm{IC}}$		-1989.9941	0.0000	0.1881	26	142
D-EM		-1989.9949	0.7136	4.9997	31,149	1
C-EM		-1989.9	-	-	2,360	-
PD	0.1	-1989.9281	0.0000	0.2521	25	638
$\mathrm{PD}^{\mathrm{IC}}$		-1989.9281	0.0000	0.2520	25	719
EM		-1989.9322	4.0901	5.0000	108,312	1
C-EM		-1989.9	-	-	1,997	-
PD	0.01	-1989.9272	0.1108	0.2270	27	943
$\mathrm{PD}^{\mathrm{IC}}$		-1989.9272	0.1108	0.2269	27	1,192
D-EM		-1989.9319	4.8230	5.0000	113,081	1
C-EM		-1989.9	-	-	1,924	-

in the mixture complexity. For this case of δ, most of the mixture probabilities are near zero; hence the algorithms must push the estimates to the boundary of the parameter space—a least favorable case for the D-EM algorithm. When the NPMLE has fewer than d support points (an overparameterized mixture problem), then the D-EM algorithm has great difficulty in allocating probability to the redundant support points. The behavior of the cumulative distribution function (CDF) of $\widehat{\mathcal{Q}}$ for the two algorithms at $\eta = 0.01$ is shown in Figure 4. It is clear that the D-EM algorithm has an extremely small step size whereas the PD has a reasonably large step size. This is due to the fact that the D-EM has retained a significantly large number of components with small jumps—an artifact of its failure to converge to the MLE in finite number of steps. From a model selection point of view, the D-EM fails to eliminate the redundant components while the PD algorithm provides a parsimonious mixture fit.

7.2 Simulation Experiment

We consider the simulated data described in Section 5.1. The data were generated from the multivariate normal mixture densities $N_p(\mathcal{Q}, \mathbf{I})$ with $p = 3$ and $n = 270$ (see Table 2) by selecting the true mixing measure for $\mathcal{Q} \in \mathcal{G}$ as the coordinates of $\boldsymbol{\theta}_j \, (j = 1, \ldots, m)$ from the set $\{-5, 0, 5\}$ in all possible combinations, with equal mass at each support vector.

Following Section 6.2, we apply the Penalized Dual algorithm in the context of building a collection of semiparametric mixture models for selected values of the sieve parameter δ. This will yield estimators with both many and few active support vectors; thereby providing a mechanism to demonstrate the superiority of our method over the D-EM algorithm across a range of applications. Both the PD and PD$^{\text{IC}}$ algorithms provide uniformly better performance, producing 6 to 40-fold improvement in CPU factor over the D-EM algorithm. As illustrated in Section 6.2, in fitting $N_p(\mathcal{Q}, \, \delta \, \boldsymbol{\Sigma})$, there is a trade-off between decrease in the sieve parameter δ and the increase in mixture complexity m; by increasing δ, we obtain a reduction in the mixture complexity m.

As discussed in Section 5.1, an important attribute of the convergence of an algorithm is the value of loglikelihood, as it indicates accuracy on a confidence interval scale. Therefore,

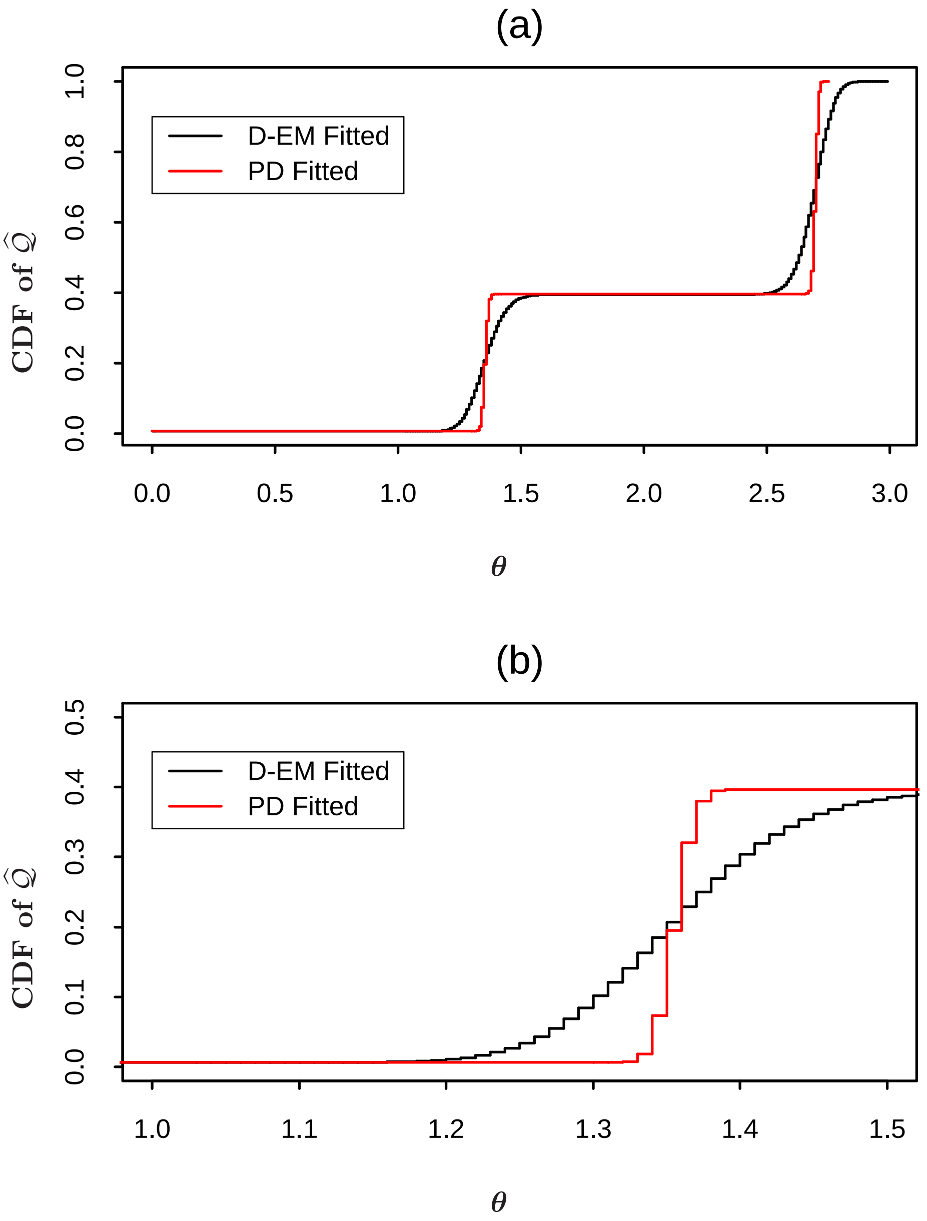

Figure 4: (a) Behavior of the cumulative distribution function (CDF) of $\widehat{\mathcal{Q}}$ for the fixed support mixture model obtained from the D-EM (black) and PD (red) algorithms at $\eta = 0.01$ for the mortality data. (b) An enlarged view of graph (a) at the first jump; the D-EM algorithm has many small jumps and retained the redundant components.

Table 4: Building a collection of semiparametric mixture models $\left\{\left(\widehat{\mathcal{Q}}_\delta,\, \delta\, \widehat{\boldsymbol{\Sigma}}\right) : \delta \in \Re_+\right\}$ using Algorithm 1 for the simulated, Fisher iris and Yeast microarray data sets. First step involved choosing the observed data matrix $\mathbf{y}$ for $\boldsymbol{\Theta}_m$ and setting $\widehat{\boldsymbol{\Sigma}} = \mathbf{S}$ in finding $l(\widehat{\boldsymbol{\pi}})$ using the PD and discrete EM (D-EM) algorithms. The estimate $\widehat{\boldsymbol{\pi}}$ obtained from the PD algorithm with the corresponding fixed support set and $\delta\,\mathbf{S}$ are employed as parameter starting values for finding $l\left(\widehat{\mathcal{Q}}_\delta, \delta\, \widehat{\boldsymbol{\Sigma}}\right)$ using the continuous EM (C-EM) algorithm.

Data	**Algorithm**			δ		
		5	2	1	0.5	0.2
Simulated	PD	-2642.8555	-2393.6817	-2313.6291	-2278.7175	-2178.5765
	D-EM	-2642.8604	-2393.6822	-2313.6299	-2278.7175	-2178.5766
	C-EM	-2313.2	-2313.2	-2192.13	-2053.37	-1876.04
Fisher Iris	PD	-629.1448	-449.8594	-376.9440	-311.5519	-192.0285
	D-EM	-629.1496	-449.8595	-376.9442	-311.5520	-192.0285
	C-EM	-379.91	-217.3	-149.63	-49.16	-136.65
Yeast Microarray	PD	-8088.9982	-5371.8998	-3691.6696	-1798.2265	-
	D-EM	-8088.9987	-5371.8999	-3691.6696	-1798.2265	-
	C-EM	-4025.3	-2626.1	-142.2	6544.0	-

in Table 4, we present the loglikelihood values obtained using various algorithms. The CPU factor for the PD algorithm over the EM algorithm ranged from ten to over forty-fold for $\delta \in \{5, 2, 1, 0.5, 0.2\}$. As predicted by the theory, for the PD and D-EM algorithms, the final $\Psi(\mathcal{Q}^{(t)})$ given by (22) does provide a guarantee on the level of algorithmic convergence $\Lambda^{(t)}$. Indeed, in some cases the bound $\Lambda^{(t)} \leq \Psi(\mathcal{Q}^{(t)})$ was very conservative. Moreover, the convergence criteria for the PD algorithm described in Section 4.1 achieved the desired accuracy in $\Lambda^{(t)}$; however, typically the PD algorithms terminated at a considerably higher accuracy than the D-EM algorithm. In order to measure this effect, we continued the D-EM algorithm to the same level of accuracy as that of the PD for $\delta = 1$. In this case, for the EM algorithm, $\mathrm{N}^{(t)} = 9,987$ at convergence, resulting in a CPU factor of 60 instead of 24.

7.3 Fisher Iris and Yeast Microarray Data Sets

We fit a mixture of multivariate normal distributions to the Fisher iris data described earlier by finding the NPMLE of $\mathcal{Q}$ and by building a collection of semiparametric mixture models for selected values of δ; results are presented in Table 4. The performance of the PD and D-EM algorithms was similar to that of the simulated data.

Instead of choosing the PD solution as the parameter starting values for the C-EM, we consider random values to demonstrate their effect on a given algorithm. It is important to recognize that the C-EM algorithm requires an a priori knowledge of m. We considered the Fisher iris data with $\delta = 1$ and randomly selected $m = 15$ data vectors from $n = 150$ as parameter starting values for the algorithm. In the ten runs of the C-EM algorithm with random starting values, $l(\widehat{\mathcal{Q}}, \widehat{\boldsymbol{\Sigma}})$ ranged from -151.99 to -179.13; all of which are sub-optimal modes due to the “poor choice” of starting values. Similar behavior of the C-EM algorithm, in reaching a sub-optimal solution, was observed by Pilla and Lindsay (2001) for the galaxy data. Without an a priori knowledge of m, choosing m can be quite a challenge for using the C-EM algorithm in large-scale practical problems.

The main technology for conducting high-throughput experiments in functional genomics is the microarray—a technical approach for assaying the abundance of mRNA for several genes simultaneously (see Hastie et al. (2001) for literature). A gene expression data set collects the expression values from a series of DNA microarray experiments with each column representing an experiment. Analysis of the expression patterns obtained from large gene arrays reveal the existence of clusters of genes with similar expression patterns. It is common to write the gene expression data of n genes, each measured at k individual array experiments (e.g., single time points or conditions) as an $n \times k$ matrix. Holter et al. (2000) analyzed a subset of the original published yeast cdc15 cell-cycle data which consist of $n = 696$ genes under $p = 12$ time points or conditions. An important scientific question is to find out which genes are most similar to each other, in terms of their expression profiles across samples. One way to organize gene expression data is to cluster genes on the basis of their expression patterns. One can think of the genes as points in $\Re^{12}$, which we want to cluster together in some fashion.

We fit multivariate normal mixtures to the yeast microarray data by finding the NPMLE of $\mathcal{Q}$. This is an example of high-dimensional modeling. We observed similar performance of the algorithms to the previous examples. Table 4 presents the loglikelihood values. Since each observation is a point in $\Re^{12}$, at $\delta = 0.2$, we obtain the empirical CDF as the MLE. That is, each observation is its own component; hence the solution is not interesting.

8 Discussion

In this article we developed a framework for approximating the continuous parameter space and created an algorithm (based on the Penalized Dual method) for finding the maximum of $l(\mathcal{Q})$; consequently an algorithm for estimating the mixture complexity. We established convergence properties of the proposed algorithm. By exploiting the inherent advantage of the penalty formulation, we derived a technique for converting the parameter estimators from the Penalized Dual problem into those for the mixture probability parameters. We established the existence of parameter estimators and derived convergence results for the Penalized Dual algorithm, for fitting overparameterized mixture models. It was shown empirically that the Penalized Dual algorithm has a faster rate of convergence, compared with the discrete EM algorithm for overparameterized mixture problems.

The algorithm based on the Penalized Dual method reaches closer to the global maximum and is robust to the choice of the support set $\mathbf{\Theta}_m \subset \Omega$ (dimensionality of the problem). These are desirable features for (1) analyzing high-dimensional data, and (2) for building a collection of semiparametric mixture models. The dimension of the dual optimization problem is fixed at d, the number of distinct observed data vectors; whereas that of the discrete EM grows with the cardinality of $\mathbf{\Theta}_m$. For discrete mixture problems, such as binomial or Poisson, often $d \ll n$; therefore, there is no dimensionality cost with the dual problem. When the cardinality of $\mathbf{\Theta}_m$ is large, the discrete EM algorithm fails to converge to the MLE, for all practical purposes, in certain mixture problems.

We derived several important structural properties of multivariate normal mixtures in which $\mathcal{Q}$ is modeled nonparametrically in the presence of an unknown variance-covariance matrix $\mathbf{\Sigma} \in \mathcal{S}$ common to all m components. The role of the sieve parameter in reducing the dimension of the mixture problem was demonstrated by creating new graphical devices, namely the Mixture Tree plots.

The proposed methods are very powerful in searching over the whole discretized parameter space and in yielding a parsimonious mixture model. The discrete EM algorithm can be very difficult, if not impossible, in yielding a parsimonious model in problems with hundreds or thousands of parameters. Such problems are becoming increasingly common due to the rapid explosion of high-throughput data in microarray data and data mining. The applications for the methods described in this article are rich. Multivariate normal mixtures arise in many different practical scenarios, including data mining, knowledge discovery, data compression, pattern recognition and pattern classification.

Appendix: Technical Derivations

A.1 Relation Between the Primal and Dual Problems

We establish the relation between the primal and dual problems at the solution using the change of variable $\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i) = (n_i/n)(w_i)^{-1}$ $(i = 1, \ldots, d)$. As a first step, we prove the following claim.

Claim. The maximization of the primal problem in (3) is equivalent to

$$\min_{\mathbf{g}_{\mathcal{Q}}} \sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)\} \tag{A.1}$$

subject to $\mathbf{g}_{\mathcal{Q}} = \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_1), \ldots, \mathrm{g}_{\mathcal{Q}}(\mathbf{y}_d)\}^T \in \Re^d_+$ and $\mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}_j) \leq 0$ for $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m$ $(j = 1, \ldots, m)$, where $\mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}_j)$ is defined in (6).

The gradient constraints $\mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}_j) \leq 0$ can equivalently be expressed as

$$\sum_{i=1}^{d} \left(\frac{n_i}{n}\right) \frac{f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)} \leq 1 \quad \text{for} \quad j = 1, \ldots, m. \tag{A.2}$$

Let $\mathcal{Q}^\star \in \mathcal{G}$ be the solution to the primal problem in (3) and let $\mathcal{Q} \in \mathcal{G}$ be any solution that satisfies constraints of the dual problem in (A.1). The equivalence between the primal problem in (3) and the dual problem in (A.1) follows by establishing that

$$\sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)\} \geq \sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i)\}. \tag{A.3}$$

Since $\log(x + \lambda) \geq \log x + \lambda/(x + \lambda)$, where $x + \lambda = \mathrm{g}_{\mathcal{Q}}$ and $x = \mathrm{g}_{\mathcal{Q}^\star}$, the above inequality yields

$$\begin{aligned} \sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)\} \quad &\geq \quad \sum_{i=1}^{d} n_i \log \{\mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i)\} + \sum_{i=1}^{d} n_i \, \frac{\{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i) - \mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i)\}}{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)} \\ &= \quad \sum_{i} n_i \log \{\mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i)\} - \sum_{i} n_i \left\{ \frac{\sum_j \pi_j \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)} - 1 \right\}. \end{aligned} \tag{A.4}$$

The second term in the right-hand side of (A.4) is less than zero since $\mathcal{D}_{\mathcal{Q}}(\boldsymbol{\theta}_j) \leq 0$ and hence the relation (A.3) holds. Therefore, the claim is established.

Define $w_i = (n_i/n)\{\mathrm{g}_{\mathcal{Q}}(\mathbf{y}_i)\}^{-1}$ so that the constraints in (A.2) become $\sum_i w_i \, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) \leq 1$

for $j = 1, \ldots, m$. From this definition of w_i, the dual problem in (A.1) can be expressed as

$$\min_{\mathbf{w} \in \Re^d_+} \left\{ \sum_{i=1}^{d} n_i \log \left(\frac{n_i}{n} \right) - \sum_{i=1}^{d} n_i \log (w_i) \right\}.$$

Equivalently, the problem is $\max\limits_{\mathbf{w}} \sum_i n_i \log (w_i)$ subject to $\mathbf{w} \in \Re^d_+$ which is the dual optimization problem in (7).

A.2 Derivation of the Penalized-Dual Estimator $\widehat{\pi}^\star_{j,\gamma}$

First, from the primal-dual relationship, it follows that

$$\widehat{w}_i = \frac{n_i}{n} \left[\sum_{j=1}^{m} \widehat{\pi}_j \, \{ f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) \} \right]^{-1} \quad \text{for} \quad \mathbf{y}_i \in \mathcal{Y};\ i = 1, \ldots, d. \tag{A.5}$$

Second, the following fixed-point equation is obtained by solving (12),

$$\widehat{w}_{i,\gamma} = \frac{n_i}{n} \left[\sum_{j=1}^{m} \left\{ p_{\boldsymbol{\theta}_j} \left(\widehat{\mathbf{w}}_\gamma \right) \right\}^{(\gamma-1)} f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) \right]^{-1} \quad \text{for} \quad i = 1, \ldots, d. \tag{A.6}$$

By comparing the right-hand sides of (A.5) and (A.6), it is clear that $\mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i)$ parallels the term $\sum_j \{p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma)\}^{(\gamma-1)} f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)$ and that the latter expression resembles a mixture density with $\{p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma)\}^{(\gamma-1)}$ playing the role of $\widehat{\pi}_j$.

A.3 Hessian Matrix of the Function $\mathcal{K}(\mathbf{z}, \gamma)$

Let $\mathbf{F} = (\mathbf{f}_{\boldsymbol{\theta}_1}, \ldots, \mathbf{f}_{\boldsymbol{\theta}_m})^T$ be an $(m \times d)$ matrix where $\mathbf{f}_{\boldsymbol{\theta}_j} = \{f_{\boldsymbol{\theta}_j}(\mathbf{y}_1), \ldots, f_{\boldsymbol{\theta}_j}(\mathbf{y}_d)\}^T$ is the d-dimensional vector. In the sequel, we denote a vector of ones by $\mathbf{1}$ (with dimension clear from the context) and the diagonal matrix with elements $\mathbf{x}$ by $\mathrm{diag}(\mathbf{x})$. Therefore,

$$\mathcal{K}(\mathbf{z}, \gamma) = \frac{1}{n} \mathbf{n}^T \cdot \mathbf{z} - \frac{1}{\gamma} \mathbf{1}^T \cdot \mathbf{p}^\gamma \quad \text{for} \quad \mathbf{z} \in \Re \text{ and } \gamma \in \Re_+,$$

where $\mathbf{n} = (n_1, \ldots, n_d)^T$ and the constraint vector

$$\mathbf{p} = \left\{ p_{\boldsymbol{\theta}_1}(\mathbf{z}), \ldots, p_{\boldsymbol{\theta}_m}(\mathbf{z}) \right\}^T$$

with $p_{\boldsymbol{\theta}_j}(\mathbf{z})$ (sometimes written as p_j for exposition) is as in (11) expressed in terms of $\mathbf{z} \in \Re$.

The Hessian matrix of $\mathcal{K}(\mathbf{z},\gamma)$ has the following elements:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{z}}\mathcal{K}(\mathbf{z},\gamma) &= \frac{\mathbf{n}}{n} - \operatorname{diag}(\mathbf{w}) \cdot \left\{ \mathbf{F}^T \cdot \mathbf{p}^{\,(\gamma-1)} \right\}, \\
\frac{\partial^2}{\partial \mathbf{z}\, \partial \mathbf{z}^T}\mathcal{K}(\mathbf{z},\gamma) &= -\operatorname{diag}(\mathbf{w} \cdot \left\{ \mathbf{F}^T \cdot \mathbf{p}^{\,(\gamma-1)} \right\} \\
&\quad -(\gamma-1)\operatorname{diag}(\mathbf{w}) \cdot \mathbf{F}^T \operatorname{diag}\left\{ \mathbf{p}^{\,(\gamma-2)} \right\}\, \mathbf{F} \cdot \operatorname{diag}(\mathbf{w}), \\
\frac{\partial}{\partial \gamma}\mathcal{K}(\mathbf{z},\gamma) &= \frac{1}{\gamma} \sum_{j=1}^{m} (p_j)^{\,\gamma} \left\{ \frac{1}{\gamma} - \log{(p_j)} \right\}, \\
\frac{\partial^2}{\partial \gamma^2}\mathcal{K}(\mathbf{z},\gamma) &= \frac{1}{\gamma}\left[-\frac{\partial \mathcal{K}(\mathbf{z},\gamma)}{\partial \gamma} + \sum_{j=1}^{m} (p_j)^{\,\gamma} \left\{ \frac{1}{\gamma} - \log{(p_j)} \right\} \log{(p_j)} - \frac{1}{\gamma^2} \right],
\end{aligned}$$

where $\mathbf{w} \in \Re^d_+$ is expressed as $\{\exp(\mathbf{z}_1), \ldots, \exp(\mathbf{z}_d)\}$ and

$$\frac{\partial^2}{\partial \mathbf{z}\, \partial \gamma}\mathcal{K}(\mathbf{z},\gamma) = -\frac{1}{\gamma} \sum_{j=1}^{m} (p_j)^{\,(\gamma-1)} \log{(p_j)}. \tag{A.7}$$

A.4 Proofs

In the sequel, we write $p_{\boldsymbol{\theta}_j}(\widehat{\mathbf{w}}_\gamma) = \widehat{p}_{j,\gamma}$ for exposition.

Proof of Theorem 2. From the fixed-point equation (A.6), we have the EM solution

$$\begin{aligned}
\widehat{\pi}_{j,\mathrm{EM}} &= \widehat{\pi}^{\dagger}_{j,\gamma} \cdot \sum_{i=1}^{d} \left(\frac{n_i}{n} \right) \frac{f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}{\sum_{k=1}^{m} \widehat{\pi}^{\dagger}_{k,\gamma}\, f_{\boldsymbol{\theta}_k}(\mathbf{y}_i)} \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m \, (j = 1, \ldots, m) \\
&= \wp_\gamma \, \{\widehat{p}_{j,\gamma}\}^{\,(\gamma-1)} \sum_i \left(\frac{n_i}{n} \right) \frac{\wp_\gamma^{-1}\, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}{\sum_k \{\widehat{p}_{k,\gamma}\}^{\,(\gamma-1)}\, f_{\boldsymbol{\theta}_k}(\mathbf{y}_i)} \qquad \text{(A.8)}
\end{aligned}$$

which follows from (16), where $\wp_\gamma$ is given in (19). From (A.6), the last equation becomes

$$\{\widehat{p}_{j,\gamma}\}^{\,(\gamma-1)} \sum_{i=1}^{d} w_i\, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i).$$

This again simplifies to $\{\widehat{p}_{j,\gamma}\}^{\,(\gamma-1)}\ \widehat{p}_{j,\gamma} = \{\widehat{p}_{j,\gamma}\}^{\,\gamma}$ due to the relationship in (15). Thus $\widehat{\pi}_{j,\mathrm{EM}} = \{\widehat{p}_{j,\gamma}\}^{\,\gamma} = \widehat{\pi}^{\star}_{j,\gamma}$ and the proof of part (a) follows. As a consequence of the EM result, the estimators are in the unit simplex $\boldsymbol{\Pi}^{\star}$ as claimed in part (b). Proof of part (c) follows by using the first inequality in part (b) in conjunction with relation (14). These two imply that $p_{\boldsymbol{\theta}}(\widehat{w}_{i,\gamma}) \leq 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$. Hence, our estimator is in the feasible region as claimed. ■

Next, we need the following lemma to prove Theorem 3.

Lemma 10 As $\gamma \to \infty$, $\wp_\gamma^{-1} \to 1$.

Proof. From the following Lyapunov's inequality (Lehmann, 1999),

$$E\Big(X^{(\gamma-1)}\Big)^{\frac{1}{(\gamma-1)}} \leq E\Big(X^{\gamma}\Big)^{\frac{1}{\gamma}},$$

one can find a bound for $\wp_\gamma$. For m number of constraints, it follows that

$$\left(\sum_{j=1}^{m} \frac{1}{m}\,\{\widehat{p}_{j,\gamma}\}^{(\gamma-1)}\right)^{\frac{1}{(\gamma-1)}} \quad \leq \quad \left(\sum_{j=1}^{m} \frac{1}{m}\,\{\widehat{p}_{j,\gamma}\}^{\gamma}\right)^{\frac{1}{\gamma}} = m^{-\frac{1}{\gamma}} \cdot 1.$$

Equivalently, $\sum_j \{\widehat{p}_{j,\gamma}\}^{(\gamma-1)} \leq m^{\frac{1}{\gamma}}$. Hence as $\gamma \to \infty$, we obtain $\wp_\gamma^{-1} \to 1$. ■

Proof of Theorem 3. From Corollary 5, we have

$$\mathcal{D}_{\widehat{\mathcal{Q}}_\gamma^\dagger}(\boldsymbol{\theta}_j) \leq \wp_\gamma - 1 \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m\,(j = 1, \ldots, m),$$

where $\mathcal{D}_{\widehat{\mathcal{Q}}_\gamma^\dagger}(\boldsymbol{\theta}_j)$ and $\wp_\gamma$ are defined in (17) and (19), respectively.

From Lemma 10, we have $\wp_\gamma^{-1} \to 1$ as $\gamma \to \infty$. Therefore, in the limit, the primal-gradient function satisfies the inequality

$$\lim_{\gamma\to\infty} \mathcal{D}_{\widehat{\mathcal{Q}}_\gamma^\dagger}(\boldsymbol{\theta}_j) \leq 0 \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m\,(j = 1, \ldots, m).$$

The compactness of the parameter space $\boldsymbol{\Pi}$ can in turn be used to establish the convergence of $\mathbf{g}_{\widehat{\mathcal{Q}}_\gamma^\dagger}$ to the maximizing value $\mathbf{g}_{\widehat{\mathcal{Q}}}$. If the vector of masses $\boldsymbol{\pi}$ for $\mathbf{g}_{\widehat{\mathcal{Q}}}$ are uniquely determined, then the masses must converge as well. This in turn implies that as $\gamma \to \infty$, the mixing distribution $\widehat{\mathcal{Q}}_\gamma^\dagger$ with $\widehat{\boldsymbol{\pi}}^\dagger$ as the vector of masses is the NPMLE. Consequently, $\widehat{\pi}_{j,\gamma}^\dagger \to \widehat{\pi}_j$ as $\gamma \to \infty$ for $j = 1, \ldots, m$. ■

Proof of Theorem 4. From (17) and the relation $\widehat{\pi}_{j,\gamma}^\dagger = \{\widehat{p}_{j,\gamma}\}^{(\gamma-1)}\wp_\gamma$, it follows that

$$\mathcal{D}_{\widehat{\mathcal{Q}}_\gamma^\dagger}(\boldsymbol{\theta}_j) = \sum_{i=1}^{d} n_i \left[\frac{\wp_\gamma^{-1} f_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}{\sum_k \{\widehat{p}_{k,\gamma}\}^{(\gamma-1)}\, f_{\boldsymbol{\theta}_k}(\mathbf{y}_i)} - 1\right] \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m.$$

From (A.6), we have $\widehat{w}_{i,\gamma} = n_i \left[\sum_k \{\widehat{p}_{k,\gamma}\}^{(\gamma-1)} f_{\boldsymbol{\theta}_k}(\mathbf{y}_i)\right]^{-1}$ and hence the last equation simplifies to

$$\wp_\gamma^{-1} \sum_{i=1}^{d} \widehat{w}_{i,\gamma}\, f_{\boldsymbol{\theta}_j}(\mathbf{y}_i) - 1 \quad \text{for} \quad \boldsymbol{\theta}_j \in \boldsymbol{\Theta}_m.$$

The desired result follows from the definition of $\widehat{p}_{j,\gamma}$. ■

Proof of Theorem 6. The proof of part (a) is a consequence of the negative definiteness of $\mathbf{H}$ and the properties of $\mathcal{P}(\mathbf{w},\gamma)$ given in Proposition 1. For part (b), let

$$\widehat{\mathbf{z}}_\gamma = \arg\max_{\mathbf{z}\in\Re} \mathcal{K}(\mathbf{z},\gamma)$$

for any fixed $\gamma \in \Re_+$. That is, $\widehat{\mathbf{z}}_\gamma$ is the maximizer of the $\mathcal{K}(\mathbf{z},\gamma)$ for a fixed γ. From equation (21) for a given $\widehat{\mathbf{z}}_\gamma$, it follows that

$$\frac{\partial}{\partial\gamma}\mathcal{K}\left(\widehat{\mathbf{z}}_\gamma,\gamma\right) = \frac{1}{\gamma}\sum_{j=1}^{m}\left\{\widehat{p}_{j,\gamma}\right\}^\gamma\left\{\frac{1}{\gamma} - \log\left(\widehat{p}_{j,\gamma}\right)\right\},$$

where $\widehat{p}_{j,\gamma}$ is expressed in terms of $\widehat{\mathbf{z}}_\gamma$. From part (c) of Theorem 2, we have $\left\{\widehat{p}_{j,\gamma}\right\}^\gamma \le 1$ for $j = 1,\ldots,m$. Therefore, $\log\left(\widehat{p}_{j,\gamma}\right) \le 0$ and hence $\partial\mathcal{K}(\widehat{\mathbf{z}}_\gamma,\gamma)/\partial\gamma > 0$ for any finite $\gamma \in \Re_+$ and fixed $\widehat{\mathbf{z}}_\gamma$. That is, for a fixed $\mathbf{z} \in \Re$, the only point at which the function $\mathcal{K}(\mathbf{z},\gamma)$ can approach its supremum is at $\gamma = \infty$. ■

Proof of Theorem 8. For exposition we drop the subscript $\mathbf{\Sigma}$ from $\mathcal{Q}^\star$ and $\widehat{\mathcal{Q}}$; however it is understood that the mixing measures have a dependence on the fixed $\mathbf{\Sigma}$.

Part (1): We first establish that $\widehat{\mathcal{Q}}$ is an NPMLE. We start with creating a path in $\mathcal{G}$ from $\widehat{\mathcal{Q}}$ to $\mathcal{Q}^\star$, by letting $\mathcal{Q}_\alpha = (1-\alpha)\,\widehat{\mathcal{Q}} + \alpha\,\mathcal{Q}^\star$ for $\alpha \in [0,1], \mathcal{Q}^\star \in \mathcal{G}$ and $\widehat{\mathcal{Q}}$ satisfying the relation (27). Note that $\mathcal{Q}_\alpha \in \mathcal{G}$; therefore, $\mathcal{G}$ is a convex set. The loglikelihood along this path satisfies

$$l\left(\mathcal{Q}_\alpha;\mathbf{\Sigma}\right) \ge (1-\alpha)\,l\left(\widehat{\mathcal{Q}};\mathbf{\Sigma}\right) + \alpha\,l\left(\mathcal{Q}^\star;\mathbf{\Sigma}\right)$$

for $\alpha \in [0,1]$ and for a fixed $\mathbf{\Sigma}$. Therefore, $l(\mathcal{Q}_\alpha;\mathbf{\Sigma})$ is a concave function for a fixed $\mathbf{\Sigma}$. The *directional derivative* of $l(\mathcal{Q}_\alpha;\mathbf{\Sigma})$ at $\mathrm{g}_{\widehat{\mathcal{Q}}}$ toward $\mathrm{g}_{\mathcal{Q}^\star}$ can be expressed, after simplification, as

$$\frac{d}{d\alpha}\,l(\mathcal{Q}_\alpha;\mathbf{\Sigma})\Big|_{\alpha=0} \quad = \quad \int \mathcal{D}_{\widehat{\mathcal{Q}}}(\boldsymbol{\mu};\mathbf{\Sigma})\,d\,\mathcal{Q}^\star(\boldsymbol{\mu}). \tag{A.9}$$

From (27), it follows that (A.9) is ≤ 0 for all $\mathcal{Q}^\star \in \mathcal{G}$ and $\widehat{\mathcal{Q}}$ satisfying (27). This result combined with the concavity of $l(\mathcal{Q}_\alpha;\mathbf{\Sigma})$ implies that $\widehat{\mathcal{Q}}$ is an NPMLE of $\mathcal{Q} \in \mathcal{G}$.

Part (2): We establish the uniqueness of the NPMLE. Suppose $\widehat{\mathcal{Q}}$ and $\mathcal{Q}^\star$ are two NPMLEs of $\mathcal{Q} \in \mathcal{G}$, then

$$l\left(\mathcal{Q}_\alpha; \boldsymbol{\Sigma}\right) = (1-\alpha)\, l\left(\widehat{\mathcal{Q}}; \boldsymbol{\Sigma}\right) + \alpha\, l\left(\mathcal{Q}^\star; \boldsymbol{\Sigma}\right)$$

for all $\alpha \in [0,1]$ and for a fixed $\boldsymbol{\Sigma}$. This implies that the derivative $d\, l(\mathcal{Q}_\alpha; \boldsymbol{\Sigma})/d\,\alpha$ at $\alpha = 0$ is exactly zero. This implies that $\mathcal{Q}^\star$ (and $\widehat{\mathcal{Q}}$) is supported on $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ [i.e., the set of zeroes of $\mathcal{D}_{\widehat{\mathcal{Q}}}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$]. Furthermore, the second derivative

$$\frac{d^2}{d\,\alpha^2}\, l(\mathcal{Q}_\alpha; \boldsymbol{\Sigma})\Big|_{\alpha=0} \quad = \quad 0,$$

which implies that

$$-\sum_{i=1}^{n} \frac{\left\{\mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i; \boldsymbol{\Sigma}) - \mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i; \boldsymbol{\Sigma})\right\}^2}{\left\{\alpha\, \mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i; \boldsymbol{\Sigma}) + (1-\alpha)\, \mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i; \boldsymbol{\Sigma})\right\}^2} = 0.$$

That is,

$$\mathrm{g}_{\mathcal{Q}^\star}(\mathbf{y}_i; \boldsymbol{\Sigma}) = \mathrm{g}_{\widehat{\mathcal{Q}}}(\mathbf{y}_i; \boldsymbol{\Sigma}) \quad \text{for all} \quad i = 1, \ldots, n.$$

However, the linear independence of the vectors $\mathbf{f}_{\boldsymbol{\mu}_j}(\mathbf{y}; \boldsymbol{\Sigma})$ for $j = 1, \ldots, K$ implies that $\{\pi_1, \ldots, \pi_K\} = \{\pi_1^\star, \ldots, \pi_K^\star\}$. That is, $\mathcal{Q}^\star = \widehat{\mathcal{Q}}$; establishing the uniqueness of the NPMLE of $\mathcal{Q}$ for a fixed $\boldsymbol{\Sigma}$. ■

References

Bickel, P., Klassen, C., Ritov, Y., and Wellner, J. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer Verlag, New York.

Charnigo, R. and Pilla, R. S. (2005). Semiparametric mixtures of generalized exponential families. *Technical Report*, Department of Statistics, Case Western Reserve University.

Connolly, A. J., Genovese, C., Moore, A. W., Nichol, R. C., Schneider, J., and Wasserman, L. (2001). Fast algorithms and efficient statistics: density estimation in large astronomical datasets. Technical report, Carnegie Mellon University, Pittsburgh, PA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. of Roy. Statist. Soc. Ser. B*, 39:1–22.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, 31:201–224.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., and Fedoroff, N. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci.*, 97:8409–8414.

Hunter, D. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32:384–406.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.*, 29:1281–1296.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics*, 10:479–484.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Stat. Assoc.*, 73:805–811.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer-Verlag, New York.

Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Stat. Assoc.*, 87:120–126.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, 11:86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part II: The exponential family. *The Annals of Statistics*, 11:783–792.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5*. Institute of Mathematical Statistics, California.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

McLachlan, G. J. and Peel, D. (2001). *Finite Mixture Models*. Wiley, New York.

Owen, A. (2001). *Empirical Likelihood. Monograph Series on Statistics and Applied Probability. Vol. 92*. Chapman & Hall, CRC Press, New York.

Pilla, R. S. and Charnigo, C. (2005). Consistent estimation and model selection in semiparametric mixtures. *Technical Report*, Department of Statistics, Case Western Reserve University.

Pilla, R. S. and Lindsay, B. G. (1996). Faster EM methods in high-dimensional finite mixtures. In *Proceedings of the Statistical Computing Section*, 166–171, Alexandria, Virginia. American Statistical Association.

Pilla, R. S. and Lindsay, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88:535–550.

Pilla, R. S. and Loader, C. (2003). The volume-of-tube formula: Perturbation tests, mixture models and scan statistics. *Technical Report*, Department of Statistics, Case Western Reserve University [E-print: arXiv:math.ST/0511503].

Renegar, J. (2001). *A Mathematical View of Interior-Point Methods in Convex Optimization*. Society for Industrial and Applied Mathematics, Philadelphia.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Amer. Stat. Assoc.*, 85:617–624.

Roeder, K. (1992). Semiparametric estimation of normal mixture densities. *Ann. Statist.*, 20:929–943.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Stat. Assoc.*, 89:487–495.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91:722–732.

Roos, C., Terlaky, T., and Vial, J.-P. (1997). *Theory and Algorithms for Linear Optimization*. Wiley, New York.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Scott, D. W. (2004a). Multivariate density estimation and visualization. In Gentle, J., Haerdle, W., and Mori, Y., editors, *Handbook of Computational Statistics: Concepts and Methds*, 517–538, New York. Springer.

Scott, D. W. (2004b). Partial mixture estimation and outlier detection in data and regression. In Hubert, M., Pison, G., Struyf, A., and Aelst, S. V., editors, *Theory and Applications of Recent Robust Methods*, 297–306, Basel. Series: Statistics for Industry and Technology, Birkhauser, Basel.

Susko, E., Kalbfleisch, J. D., and Chen, J. (1999). Computational methods for mixture estimation. In *Proceedings of the Interface: Models, Predictions and Computing*, (K. Berk and M. Pourhmadi, ed.) 432–438, Vol. 31.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103.